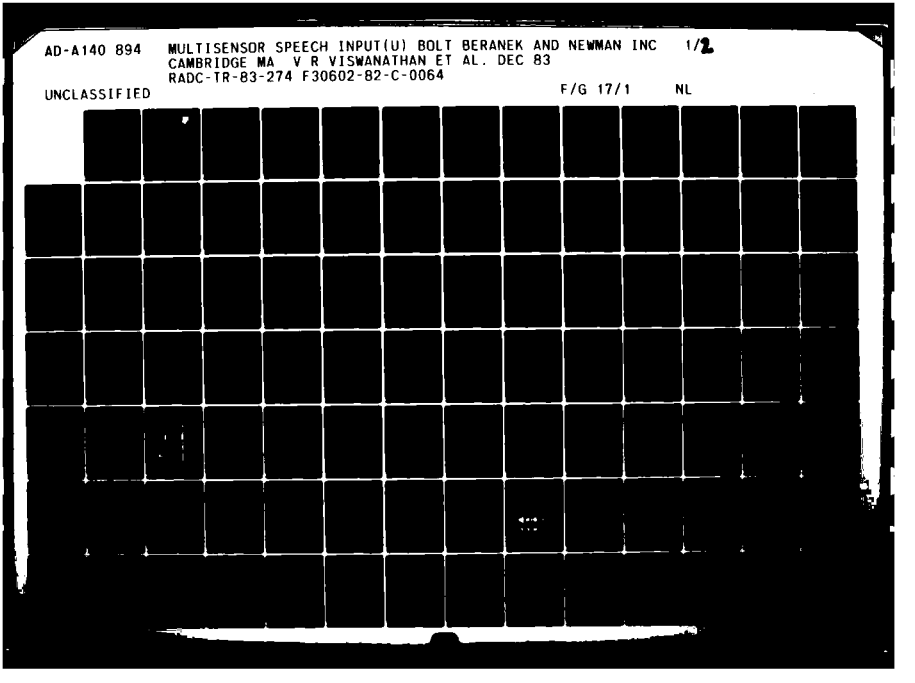
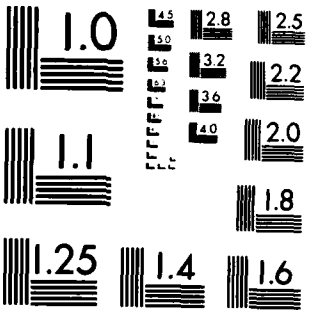
AD-A140 894 MULTISENSOR SPEECH INPUT(U) BOLT BERANEK AND NEWMAN INC 1/2
CAMBRIDGE MA V R VISWANATHAN ET AL. DEC 83
RADC-TR-83-274 F30602-82-C-0064

UNCLASSIFIED F/G 17/1 NL





(12)

RADC-TR-83-274
Final Technical Report
December 1983

MULTISENSOR SPEECH INPUT

Bolt, Beranek and Newman, Inc.

Vishu R. Viswanathan, Kenneth F. Karnofsky, Kenneth N. Stevens
and M. N. Alakel

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

This effort was funded totally by the Laboratory Directors' Fund

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-83-274 has been reviewed and is approved for publication.

APPROVED: William I. Lundgren

WILLIAM I. LUNDGREN, Capt, USAF
Project Engineer

APPROVED: Thadeus J. Domurat

THADEUS J. DOMURAT
Acting Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: John A. Ritz

JOHN A. RITZ
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-83-274	2. GOVT ACCESSION NO. A140894	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) MULTISENSOR SPEECH INPUT		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report February 82 - July 83
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Vishu R. Viswanathan M. N. Alakel Kenneth F. Karnofsky Kenneth N. Stevens		8. CONTRACT OR GRANT NUMBER(s) F30602-82-C-0064
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bolt Beranek and Newman, Inc. 10 Moulton Street Cambridge MA 02238		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61101F LDFP02C2
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAA) Griffiss AFB NY 13441		12. REPORT DATE December 1983
		13. NUMBER OF PAGES 114
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: William I. Lundgren, Capt, USAF (IRAA)

This effort was funded totally by the Laboratory Directors' Fund.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multisensor
Speech Enhancement
Noise Reduction

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The use of multiple sensors to transduce speech was investigated. A data base of speech and noise was collected from a number of transducers located on and around the head of the speaker. The transducers included pressure, first order gradient, second order gradient microphones and an accelerometer. The effort analyzed this data and evaluated the performance of a multiple sensor configuration. The conclusion was: multiple transducer configurations can provide a signal containing more useable

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

speech information than that provided by a microphone.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

The authors would like to thank Michael Cowing (Department of Defense) for providing a EV 985 microphone and Anton Segota (RADC/EEV) and Lee Kline (NRL) for providing a Vought second-order gradient microphone. The two prototype microphones were used in the sound-field measurements gathered in this project.

1. INTRODUCTION

The overall objective of this basic research project was to study, investigate, and develop multisensor configurations of existing sensors with the general goal of providing better transduction of speech signals than is currently possible with single microphones. In this section, we state the specific goals of this research, describe our measurement-based approach, and present the highlights of the work.

1.1 Research Goals and a Measurement-Based Approach

The multisensor configurations that we have sought to develop must provide additional, new acoustic information not presently used, a more reliable and accurate transduction of the presently used information, or a more robust way of extracting the same information in a hostile acoustic environment. The transduced signals of the multiple sensors can be processed to provide either a single speech signal by combining them or several parallel or independent inputs. Such multisensor speech input will undoubtedly play an important role in the design of robust and high-performance speech transmission, speech coding, and speech recognition systems. It is also conceivable that the multisensor speech input, especially the parallel-input version, may lead to the development of new speech processing methodologies.

In our work, we have employed a measurement-based approach. In this approach we have performed detailed measurements of the sound field near the lips using pressure and pressure gradient microphones and of the vibrations on the head and neck of the talker using surface transducers (accelerometers). The amplitude and spectral properties of the sound field in the vicinity of the talker are known at present only in a qualitative manner and as predicted from ideal models. One of the goals of this work has been to conduct an experimental study to map out the properties of the sound field quantitatively for a number of different classes of speech sounds in order to provide a rational basis for the selection or design of multisensor configurations for transducing the speech signal with desired properties. Another important aspect of our work is the computer analysis of the measured sound-field data to investigate the long-term and short-term spectral properties of the speech signal transduced by different sensor types and locations and to calculate, as a measure of speech intelligibility, the so-called articulation index, assuming cockpit noise typical in fighter aircraft.

1.2 Highlights of the Work

Given below is a list of the highlights of our work:

- Design, fabrication, and calibration of a helmet-mounted array of three closely-spaced electret microphones, to measure sound pressure, first-order pressure gradient, and second-order pressure gradient (Section 2).

- Detailed measurements of the sound field in the vicinity of each of five talkers and of the vibrations on the head and neck of the talker, as he or she read a specially selected set of speech materials in a noise-free anechoic chamber. We used the three-microphone array, a standard condenser microphone, a prototype first-order gradient microphone, a prototype second-order gradient microphone, and a miniature, lightweight accelerometer, in making these measurements at each of eleven locations (Section 3).

- Computer analysis of the long-term and short-term spectral amplitudes for the measured sound-field data (Section 4).

- Computation of the articulation index for different sensor types and positions, assuming ambient noise typical in fighter aircraft cockpit (Section 4).

- Determination of the best location for each sensor, from the results of the long-term spectral analysis and the articulation index scores. This aspect is particularly important because of the recent widespread use of close-talking microphones in a variety of speech processing applications (Section 4).

- Development and investigation of a two-sensor configuration consisting of an accelerometer and a gradient microphone. The two-sensor signal is a weighted sum of filtered versions of the accelerometer signal and the gradient microphone signal (Section 5).

- Intelligibility and quality evaluation of speech transduced by various single sensors and two-sensor configurations under three conditions of aircraft cockpit noise. The results of this evaluation show clearly that a given two-sensor configuration produces higher speech intelligibility and quality than does its constituent gradient microphone. The performance improvement provided by the two-sensor signal generally increases with the overall noise level (Section 6).

2. DESIGN, FABRICATION, AND CALIBRATION OF A THREE-MICROPHONE ARRAY

As mentioned in Section 1.1, our approach to the problem of multisensor speech input requires making detailed measurements of the sound field near the talker's lips during speech production. To measure sound pressure, first-order pressure gradient, and second-order pressure gradient, we designed, fabricated, and calibrated an array of three small closely-spaced electret pressure microphones. The advantages of using this array instead of three separate pressure and pressure-gradient microphones are: 1) the array can be easily located in any desired position and orientation, and 2) measurements of sound pressure and pressure gradients can be performed simultaneously. With three separate microphones (each having its own windscreen), we may not be able to use them simultaneously, without interference, especially for positions very close to the mouth.

If the three electret microphones used in the array, denoted by M1, M2, and M3 (M1 being closest to the mouth and M3 being the farthest), had identical amplitude and phase characteristics, then the first difference in the microphone signals (M1-M2 or M2-M3) would be a reasonable estimate of the first-order pressure gradient, and the second difference (M1-2M2+M3) would be a reasonable estimate of the second-order pressure gradient. However, commercially available electret microphones exhibit some amplitude and phase differences between different units. These

differences must be adequately compensated for, before we calculate the first and second differences in the microphone signals. Below, we first describe a helmet-mounted boom for the three microphone array and then present in detail the design, fabrication, and calibration of the array.

2.1 A Helmet-Mounted Boom

We designed and built a helmet-mounted boom for the three-microphone array. One end of a metal rod is attached to the top of the helmet, and the other end has an adjustable clamp to hold the microphone array. The metal rod can be moved front-to-back or sideways by the adjustable knob at the top of the helmet. The clamp holding the microphone array has provisions so that the array can be moved up or down and closer to or away from the talker. Therefore, with this boom, it is quite convenient and simple to place the array at any position and orientation within a few centimeters from the lips of the talker. Also, as the helmet is worn by the talker, the positioning of the array relative to his mouth is preserved should the talker move his head slightly during measurements.

2.2 Design and Fabrication of a Three-Microphone Array

Figure 1 shows top and side views of the three-microphone array designed and fabricated. The microphone assembly consists of a calibrated, hollow metal rod,

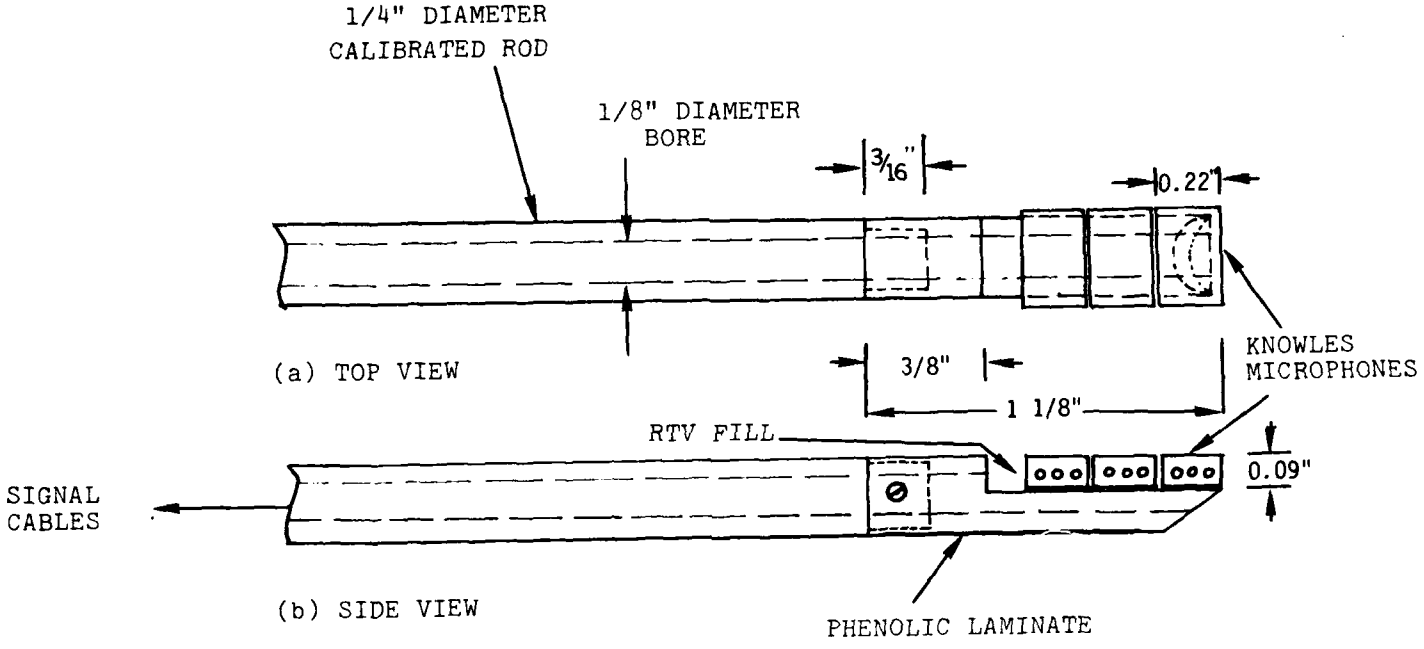


FIG. 1. Top and side views of the three-microphone assembly.

with its one end attached (with a screw) to a piece of laminated phenolic material that holds three closely-spaced Knowles electret microphones (model number BT-1759). The microphones are mounted with epoxy. Three cables, each 2 feet long, are connected to the microphone leads, run through the bore hole in the calibrated rod, and attached to a 5-pin connector. The microphone array is attached to the helmet-mounted boom with an adjustable clamp arrangement.

2.3 Amplitude and Phase Measurements of the Three Electret Microphones

Before we fabricated the three-microphone array, we picked two samples of the Knowles BT-1759 electret microphone and performed amplitude and phase measurements to determine the differences between the two samples and the type of compensation required to minimize the observed differences. After the three-microphone array was fabricated, we repeated the amplitude and phase measurements on the three microphones in the array. Since the results of the two sets of measurements are essentially the same, we report below on the measurements we performed on the microphone array.

As a reference standard for the measurements, we used a pistonphone-calibrated Bruel and Kjaer 1/2-inch condenser microphone. We performed the amplitude and phase measurements in an anechoic chamber for frequencies above 400 Hz and in a traveling wave tube for frequencies from 50 to 400 Hz.

The amplitude responses of the three microphones M1, M2, and M3 were found to be uniform within ± 1.5 dB from 80 Hz to 11 kHz, and they tracked each other reasonably well.

For making the phase measurements, we oriented the microphone array in the direction of the sound source (M1 closest to the source) so that the sound wave was at grazing incidence and was traveling along the microphone array. For this setup, given that the center-to-center microphone spacing is 0.22 in, the theoretical phase difference (caused by this spacing) between M1 and M2 or between M3 and M2 is given by:

$$\theta = \frac{0.22}{12 \times 1130} 360\ f = 0.00584\ f, \qquad (1)$$

where we have assumed the sound velocity to be 1130 feet/s, and f is the frequency in Hz. From the measured phase responses of the three microphones, we computed the phase differences between M1 and M2 and between M3 and M2, and compared them against the theoretical phase differences given by (1).

Considering first the anechoic chamber measurements for frequencies above 400 Hz, we found that both M1-M2 and M3-M2 phase differences tracked the theoretical phase difference reasonably well, with the average error over 400 Hz-5 kHz being about 1.0 degree for M1-M2 and about 1.5 degrees for M3-M2. Both cases exhibited anomalous phase differences at about 4 kHz and for frequencies above 6 kHz; these may be due to local reflections and scattering.

Let us now consider the low-frequency phase measurements that we performed in the traveling wave tube. For the frequencies 100, 200, 300, and 400 Hz, the M1-M2 phase differences were 13.5, 7.5, 4.0, and 2.5 degrees, respectively, and the M3-M2 phase differences were 4.5, 4.0, 3.5, and 3.5 degrees, respectively. We note that the theoretical phase differences at these frequencies are about 0.6, 1.2, 1.8, and 2.4 degrees, respectively. The relatively large deviations between the measured and the theoretical phase differences may be due to differences in internal impedances of the individual microphones. Clearly, the deviations are much worse for M1-M2 than for M3-M2. Further, the amplitude response at 100 Hz was +2.2 dB for M1, +0.5 dB for M2, and -0.8 dB for M3, relative to the response of the standard condenser microphone. Based on these considerations, we decided to replace M1 with another sample of the Knowles electret microphone. For this purpose, we measured the amplitude response of each of two sample units we had and chose the one with a better amplitude response, especially at the low frequencies.

As for the large phase differences at the low frequencies, we decided to compensate them electronically by including R-C (resistor-capacitor) networks inside the conditioner box, which is described next.

2.4 Conditioner Box

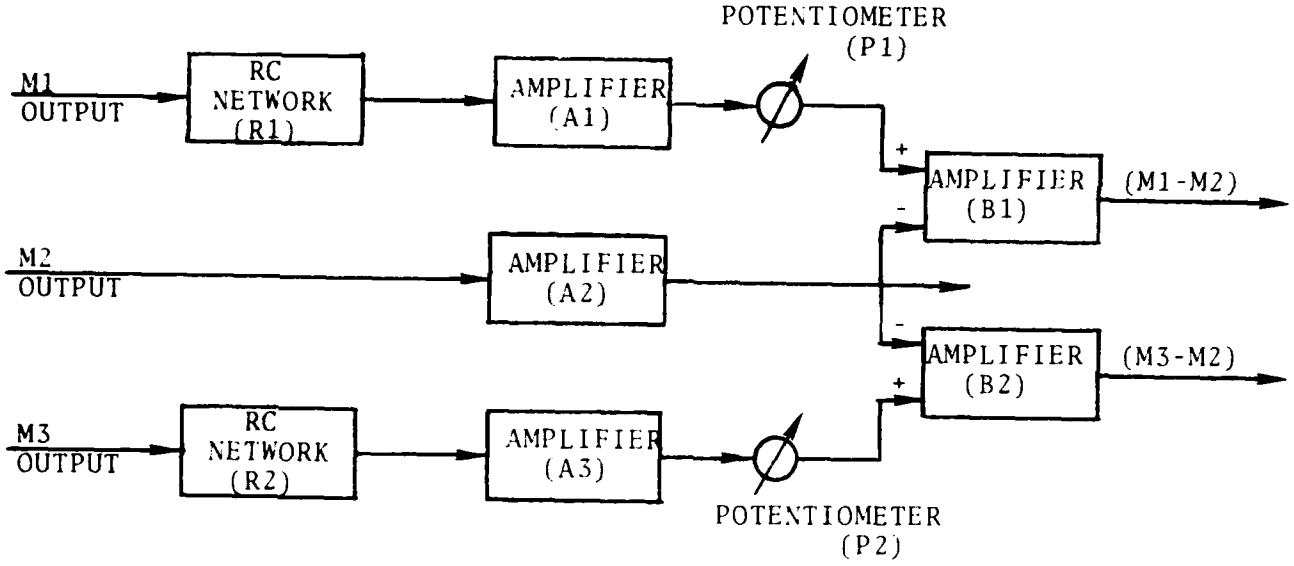
The purposes of the conditioner box are 1) to include provisions for phase and amplitude compensations of the microphone signals, 2) to provide sufficient amplification of the microphone signals, and 3) to calculate the first and second differences of the microphone signals. Figure 2 shows a block diagram of the circuitry we fabricated as the conditioner box. The conditioner has three channels with inputs provided by the three microphones M1, M2, and M3. The RC networks shown at the top and bottom channels are used to compensate for the large phase differences at the low frequencies between M1 and M2 and between M3 and M2. Each RC network has a fixed capacitor (0.01 μF), a fixed resistor (177K Ohms for the top channel and 98K Ohms for the bottom channel), and a potentiometer (50K Ohms). Let us denote the adjustable resistors of the two RC networks as R1 (top channel) and R2 (bottom channel). As shown in Fig. 2(a), the conditioner box has two stages of amplifiers. At the first stage, each channel has a buffer-amplifier with a gain of about 20 dB. For the first-stage amplifiers A1, A2, and A3, we used the Motorola operational amplifier Model MC1741CP. The second stage consists of two differencing amplifiers B1 and B2, each with a gain of about 10 dB. B1 produces the difference M1-M2, and B2 produces M3-M2. The potentiometers P1 and P2 at the input of the amplifiers B1 and B2, respectively, are used to compensate for the sensitivity differences between the respective

microphones. For the amplifiers B1 and B2, we used the low-noise operational amplifier Signetics Model NE5534AN. Either (M1-M2) or (M3-M2) can be used as the first-order pressure gradient.

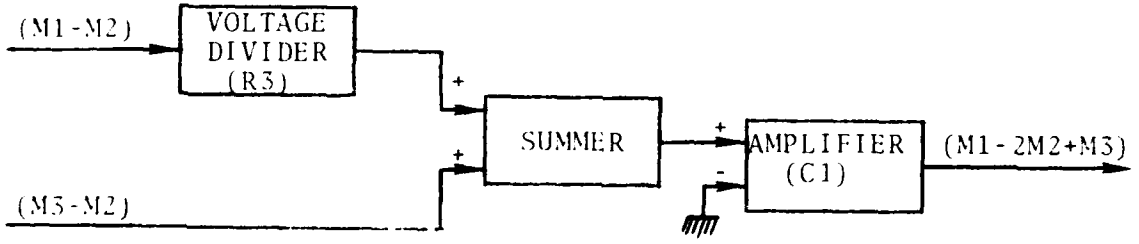
Figure 2(b) shows the part of the circuit that calculates the second-order gradient (M1-2M2+M3) from the two first-order gradients (M1-M2) and (M3-M2). The purpose of the 5K-Ohm potentiometer R3 (voltage divider) at the amplifier input is to compensate for the amplitude differences of the first-order gradient inputs (M1-M2) and (M3-M2). The potentiometer was located at the (M1-M2) input since its amplitude was found to be higher than the amplitude of the (M3-M2) input. For the amplifier, we used the Motorola operational amplifier Model MC1741CP. We included a toggle switch around the amplifier to select either about unity gain (switch closed) or about 20 dB gain (switch open).

We installed a 25-foot extension cable with appropriate connectors at its ends, to allow us to connect the microphone outputs to the conditioner box for subsequent calibration tests in the anechoic chamber and in the traveling wave tube.

Next, we present a theoretical analysis of the accuracy required in compensating for the phase and amplitude differences between the three microphones used in the array.



(a) First-order pressure gradient (M1-M2) or (M3-M2).



(b) Second-order pressure gradient (M1-2M2+M3).

FIG. 2. A block-diagram description of the circuitry used in the conditioner box.

2.5 Theoretical Analysis for Phase and Amplitude Compensation

If two adjacent microphones have phase characteristics that differ by just a few degrees at low frequencies, there will be a significant error in the estimate of pressure gradient that is obtained simply by taking the difference in the outputs of the two microphones. Thus, we need to obtain an estimate of the precision with which it is necessary to match the amplitude and phase characteristics of the microphones in the array.

Let us suppose that two microphones are located at distances r_1 and r_2 from a simple source, and that the sound pressures sensed at these locations are p_1 and p_2. We assume that the microphones are closely spaced, so that $r_2 - r_1$ is small compared to a wavelength. The sound pressures p_1 and p_2 at the two microphones can be written

$$p_1 = \frac{A}{r_1} e^{-jkr_1}$$
$$p_2 = \frac{A}{r_2} e^{-jkr_2} \qquad (2)$$

where A is a constant, $k=2\pi f/c$, f=frequency, and c=velocity of sound. The pressure difference is given by

$$p_1 - p_2 = \frac{A}{r_1} e^{-jkr_1} \left[1 - \frac{r_1}{r_2} e^{-jk(r_2-r_1)} \right] \qquad (3)$$

$$\cong \frac{A}{r_1} e^{-jkr_1} \left[1 - \frac{r_1}{r_2} \right], \qquad (4)$$

if $k(r_2 - r_1) \ll \lambda/2$ i.e., if $r_2 - r_1 \ll \lambda/4$, where λ is a wavelength. We can write the ratio

$$\frac{p_1 - p_2}{p_1} \cong 1 - \frac{r_1}{r_2} . \qquad (5)$$

For the microphone configuration we have fabricated, the spacing between the microphones is 0.56 cm. Thus we have

$$[20 \log |(p_1 - p_2)/p_1|] \cong 20 \log[0.56/(r_1 + 0.56)] . \qquad (6)$$

This ratio (in decibels) is plotted as a function of r_1 in Fig. 3. For ideal microphones whose output voltages are exactly proportional to the applied sound pressures, this curve indicates how the difference in output voltages varies with distance from a simple source.

Let us assume that the voltage outputs from the two microphones are v_1 and v_2, and that the difference between these voltages is detected. When calibrating the two microphones to determine the extent to which they deviate from having identical characteristics, a simple procedure to use is to place identical sound pressures at the two microphones and to measure the ratio $20 \log |(v_1 - v_2)/v_1|$. As a rule of thumb, this ratio, which can be called the common mode rejection, should be at least 10 dB less than the ratios plotted in Fig. 3 if the voltage difference $v_1 - v_2$ is to provide a sufficiently accurate measure of the

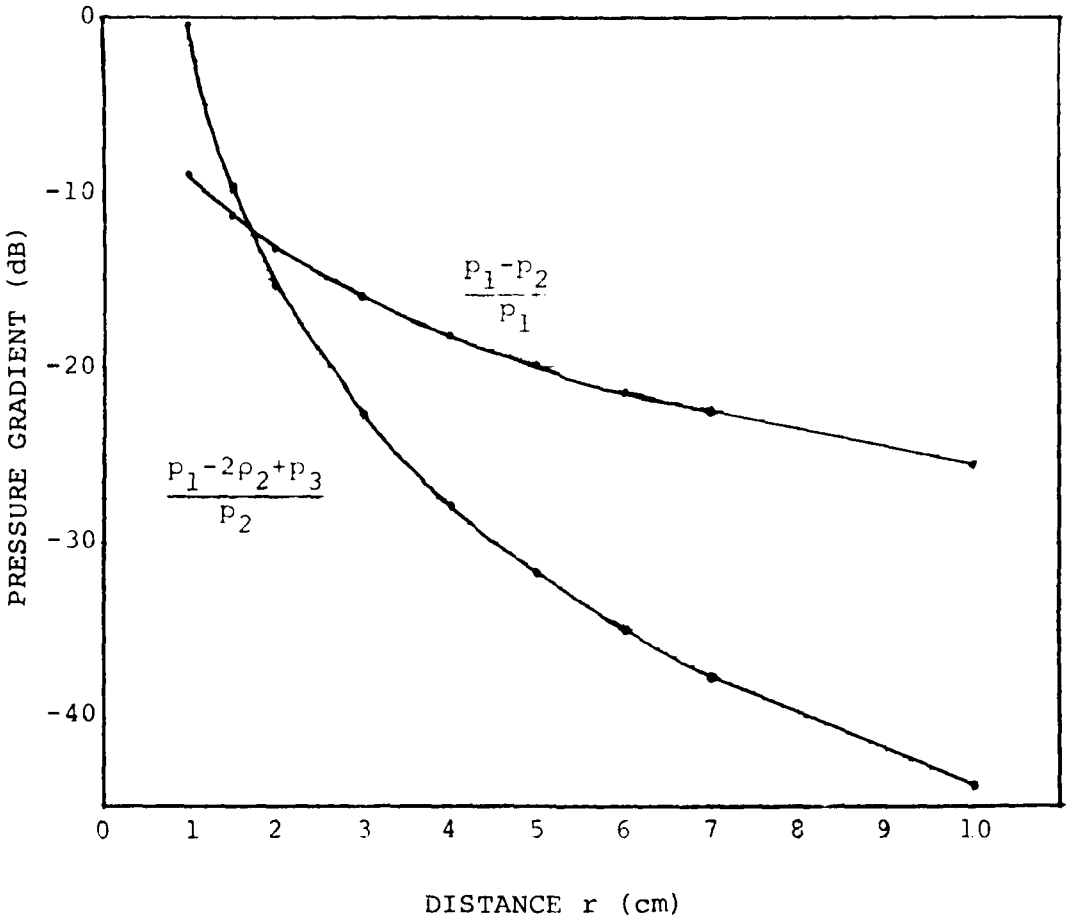


FIG. 3. Output of ideal first-order and second-order gradient microphones at various distances from a simple source, expressed as a ratio of the output of one of the microphones in the array. The distance r is r_1 for the first-order gradient case and r_2 for the second-order gradient case.

pressure gradient when the two microphones are located near a simple source. Thus, for example, if the microphones are adjusted to give a common mode rejection of -32 dB, then the voltage difference provides a good measure of the pressure gradient for distance r_1 out to about 6 cm. Since we are primarily interested in distances less than this value, a common mode rejection of -32 dB down to the lowest frequency of interest (100-200 Hz) is sufficient.

A similar analysis can be carried out for a three-microphone array arranged to measure pressures p_1, p_2, and p_3 at distances r_1, r_2, and r_3 from a simple source. The second-order gradient in sound pressure, relative to pressure p_2, can be written

$$\frac{p_1-2p_2+p_3}{p_2} = \frac{r_2}{r_1} e^{-jk(r_1-r_2)}-2 + \frac{r_2}{r_3} e^{-jk(r_3-r_1)} \qquad (7)$$

If we assume $r_1-r_2 = r_2-r_3$, and $r_1-r_2 << \lambda/4$, we obtain

$$\frac{p_1-2p_2+p_3}{p_2} \cong 2 - r_2\left(\frac{1}{r_1} + \frac{1}{r_3}\right). \qquad (8)$$

For $r_1-r_2 = 0.56$ cm, this ratio, in dB, is plotted as a function of r_2 in Fig. 3.

If we again use a criterion that the electrical outputs of the microphones should be compensated to give a common mode rejection that is at least 10 dB below the values given on this curve, then we would need a common mode rejection of about 45 dB

in order to obtain a sufficiently accurate measure of the second-order pressure gradient out to a distance of r_2=6 cm.

2.6 Phase and Amplitude Compensation

To compensate for the differences in the amplitude and phase responses of the three microphones and hence to obtain good estimates of the first-order and the second-order pressure gradients, we placed the three-microphone array in the traveling wave tube, applied identical sound pressures at the microphones, and adjusted the resistors P1, P2, R1, R2, and R3 in the conditioner box (see Figure 2) as described below. If the three microphones were identical, we would get zero volts at the outputs corresponding to the first-order gradients (M1-M2) and (M3-M2) and to the second-order gradient (M1-2M2+M3). Since the microphones are not identical, these outputs will be nonzero. Denoting the output voltages of the microphones as v_i, i=1,2,3, we define three common mode rejection (CMR) functions, as follows:

$$CMR12 = 20 \log_{10} |(v_1 - v_2)/v_1|$$

$$CMR32 = 20 \log_{10} |(v_3 - v_2)/v_2| \qquad (9)$$

$$CMR123 = 20 \log_{10} |(v_1 - 2v_2 + v_3)/v_2|$$

As we noted above in Section 2.5, the CMR's should be at least 10 dB less than the respective theoretical pressure ratios plotted in Fig. 3. We optimized (i.e., minimized) CMR12 by adjusting the resistors R1 and P1, and CMR32 by adjusting R2 and P2. Since we used a pure-tone signal in the traveling wave tube, another variable of interest is the frequency at which this optimization is performed. We considered frequencies in the range 100-400 Hz, since our measurements (Section 2.3) did not indicate any significant problems at higher frequencies.

We optimized CMR12 as follows. We observed the output (M1-M2) from the amplifier B1 (see Fig. 2(a)) on an oscilloscope and using an rms voltmeter. We adjusted P1 so that this output reached a minimum. The outputs of the buffer amplifiers A1 and A2 were then displayed in the x-y mode on another oscilloscope, with A1 output connected to the y-input channel and A2 output connected to the x-input channel. The resulting display, known as Lissajous pattern, will be a 45-degree straight line if the input signals are in phase and an ellipse if the signals are out of phase with each other. We adjusted R2 so that the display was approximately a 45-degree straight line. We then readjusted P1 to obtain a minimum reading on the voltmeter. The back and forth

adjustments of P1 and R1 were continued until no significant changes occurred in the Lissajous pattern and in the voltmeter reading. We followed the same procedure to optimize the P2 and R2 settings associated with the microphone pair (M3,M2).

We performed the above optimization at each of the three frequencies 100, 200, and 300 Hz, and then measured CMR12 and CMR32 as a function of frequency over the range 100-400 Hz. Clearly, the measured CMR's were minimum at the optimization frequency and increased progressively as we moved away from this frequency. The CMR responses were better for the 200-Hz and 300-Hz optimizations than for the 100-Hz optimization. Of the 200-Hz and 300-Hz optimizations, we found that the first one produced better CMR values at 100 Hz and 200 Hz, and the second one produced better values at 300 Hz and 400 Hz. As a compromise between the two responses and to get a good overall response, we repeated the optimization at 250 Hz, and measured the resulting responses over 100-400 Hz at 50 Hz intervals.

From the measured responses and from Fig. 3, we expect that, for frequencies above 200 Hz, the measured first-order gradient estimate (M1-M2 or M3-M2) will be good for distances of up to about 6 cm from the sound source.

To determine the variation over time of the optimized settings of the resistors, we repeated three times, over a span of one week, the optimization for the two first-order gradients

(M1-M2) and (M3-M2) at 250 Hz. As expected, the greatest changes occurred at the optimization frequency of 250 Hz, and these were 2.5 dB for CMR12 and 4.5 dB for CMR32. Changes at other frequencies were small (0.3 - 1.0 dB). Since the larger changes occurred at frequencies where the CMR was well below the level we require for accurate estimation of the gradient, we do not anticipate any problems caused by drifts in optimum settings.

Proper adjustment of the potentiometer R3 of the second-order gradient circuit (see Fig. 2(b)) could not be made in the traveling-wave tube, since the two first-order gradients were nearly zero at the optimized condition and the overall output (M1-2M2+M3) remained nearly constant (approximately zero) irrespective of the setting of R3. To optimize the CMR function CMR123 by adjusting R3, we require ideally a sound field that produces equal but non-zero values for (M1-M2) and (M3-M2). For this purpose, we used the Bruel and Kjaer artificial voice (Model number 4215) as sound source in our anechoic chamber. The three-microphone array was set up in front of the source, so that it was in alignment with the center of the source, with M1 being closest to the source. The microphone array was first located near the "lips" of the source and then moved gradually away from the source until the two quantities (M1-M2) and (M3-M2) were very nearly equal. We then adjusted the potentiometer R3 so that the output (M1-2M2+M3) or equivalently the CMR123 function was minimized. Since the anechoic chamber introduces distortions for

frequencies below 400 Hz, we adjusted R3 at 400 Hz and checked the overall output for several frequencies above 400 Hz.

2.7 Calibration Measurements with the B & K Artificial Voice

Having compensated for the inter-microphone differences in phase and amplitude responses, we made a series of sound-field measurements in the anechoic chamber, using the B & K artificial voice (Model No. 4215) as source and using the three-microphone array. The purpose of this measurement task was to examine if the three-microphone array yields reasonable estimates of the first-order and the second-order pressure gradients. The B & K artificial voice is a well-defined sound source and covers the frequency and pressure range that is normally covered by the human voice. It is of cylindrical shape and consists of a small loudspeaker mounted in a metal housing. We suspended the artificial voice and the three-microphone array from the ceiling of the anechoic chamber, using twine and tape. The microphone array was set up in front of the artificial voice source and in alignment with its center, with M1 being closest to the source. The location of the microphone array was measured as the distance of M1 from the lips of the artificial voice source. This distance was varied from 0 to 10 cm, in steps of 1 cm. The measurement setup included several instruments outside the anechoic chamber, such as B & K Beat Frequency Oscillator Model 1022 and B & K Graphic Level Recorder Model 2305, and a standard

condenser microphone (B & K Model 4133) located inside the chamber at 30 cm from the voice source. For each distance setting of the microphone array, we obtained the strip-chart plots of the frequency responses (over 0-7 kHz) of M1, M2, M3, (M1-M2), (M3-M2), and (M1-2M2+M3). For the same distance settings, we also measured the frequency responses of the output of a Shure SM-12 noise-cancelling microphone.

From the strip-chart plots, we tabulated, for each distance setting and for each of the frequencies 400, 500, 1000, 2000, 3000, and 5000 Hz, the values in dB of M1, M2, M3, (M1-M2), (M3-M2), (M1-2M2+M3), SM-12 output, M1/(M1-M2), and M2/(M1-2M2+M3). We also computed the theoretical values of the last two quantities, using the exact formulas (3) and (7) given in Section 2.5.

For each frequency, we then plotted, as a function of the distance of M1 from the source, three sets of graphs: 1) M1, (M1-M2), and (M1-2M2+M3) vs. distance; 2) (M1-M2) and SM-12 output vs. distance; and 3) theoretical and measured values of M1/(M1-M2) and M2/(M1-2M2+M3) vs. distance.

We examined in detail the various measurements and the plots. We summarize our findings as follows. First, the first-order gradient (M1-M2) tracks the output of SM-12 very well at all frequencies, indicating that (M1-M2) should have about the same noise-cancelling property as that of SM-12. Second, the

measured values of M1/(M1-M2) track the theoretical values well only for distances of 2 cm and above. The main reason, we believe, is that the theoretical values were computed under the assumption of a simple source (i.e., 6 dB drop in the sound pressure for every doubling of the distance) and that this assumption is not valid for small distances. Third, there are large discrepancies between the theoretical and the measured values of M2/(M1-2M2+M3), particularly at distances over 6 cm for low frequencies and over 2 or 3 cm for high frequencies. Specifically, the plot of the measured value vs. distance reaches a plateau much sooner than the theoretical plot does. Also, whereas the theory indicates that the second-order gradient falls off with distance much more rapidly than does the first-order gradient, the measurements we made do not show that result.

In one of our attempts to understand the problem with our second-order gradient estimate, the B & K voice source was damaged. Rather than getting the B & K source repaired, we decided to construct a sound source that is approximately a simple, point source. Besides the fact that the theoretical pressure gradient expressions against which we were comparing the measured values are valid only for a simple source, a simple, point source provides an effective way of optimizing the potentiometer R3 for the second-order gradient (see below).

2.8 Calibration Measurements with a Simple Source

The sound source we constructed consists of a 3-Watt speaker enclosed in a cylindrical box and a long tube with a small diameter, which is fitted into an opening in the box. The inner and outer diameters of the tube are, respectively, 0.225 and 0.312 inches. In our experiments, we used one of two tubes: a short tube 8 15/16 inches in length and a long tube of 13 7/16 inches. Ideally, we would like the sound output from this source to come from the opening of the tube only. To minimize the sound radiation from the surfaces of the box and the tube, we sealed these surfaces using duct sealer. To obtain reasonably large sound pressures from this source, we needed to tune the frequency of the source to be one of the resonances of the tube. The resonant frequencies at which we made sound-field measurements are: 790 Hz, 1.61 kHz, 2.35 kHz, and 3.15 kHz for the short tube. We used the long tube at only one resonant frequency: 465 Hz.

In our experiments, we observed that as we progressively moved the microphone array away from the source (5 cm or farther), the measured values were increasingly noisy. This might be one of the factors responsible for the substantial difference, observed at large distances, between the measured and the theoretical values of the second-order gradient. To improve the signal-to-noise ratio of the measurements, we decided to filter the outputs from the conditioner box using a 3rd-octave

filter with its center frequency approximately equal to the frequency of the sound source. For this purpose, we used a 3rd-octave filter bank and selected the filter that produced the largest output.

In an effort to improve our second-order gradient estimate, we decided to reoptimize the setting of the potentiometer R3 (see Fig. 2(b)). For this step, recall that we require ideally a sound field that produces equal but non-zero values for (M1-M2) and (M3-M2). Since the sound source we constructed is approximately a point source, the above condition can be achieved by placing the microphone array vertically and close to the source so that M2 is in alignment with the axis of the tube and M1 and M3 are situated symmetrically on either side of this axis. The second-order gradient output is then (M3-M2) + k(M1-M2), where k is a positive fraction that is less than 1 and whose value depends upon the setting of R3. We placed the microphone array vertically, as mentioned above, at about 1 cm from the source, and adjusted R3 so that the second-order gradient output was 6 dB above the output (M3-M2) i.e., k(M1-M2) equals (M3-M2).

We then set up the microphone array in front of the simple source, so that the array was in alignment with the axis of the tube. The location of the microphone array was measured as the distance of M1 from the tube. We varied this distance from 1 to 10 cm, in steps of 1 cm. (The actual distances of M1 were: 1.28,2.28,...,10.28 cm, respectively, where the distance is from

the tube to the center of M1.) For each distance setting, we measured M1, M1-M2, M3-M2, and M1-2M2+M3 at each of the frequencies 465 Hz, 790 Hz, 1.61 kHz, 2.35 kHz, and 3.15 kHz. A B & K condenser microphone was set up at a distance of 30 cm from the source as a standard microphone. We adjusted the source level at each frequency to get a constant level of 74 dB at the standard microphone.

From the measured data, we computed the two quantities (in dB) M1/(M1-M2) and M1/(M1-2M2+M3), and compared them against their theoretical values, which we computed using the exact formulas given in Section 2.5. The measured function of M1/(M1-M2) vs. distance tracks well the theoretical function, at each of the five frequencies. The agreement between the measured and calculated values of M1/(M1-2M2+M3), however, is good only up to about 4 cm. Plots of the theoretical and measured values at 465 Hz are given in Fig. 4 for M1/(M1-M2) and in Fig. 5 for M1/(M1-2M2+M3). Since the amplification gains introduced by the conditioner box make the comparison of the absolute values inappropriate, we have moved the plot of the theoretical values down by about 32 dB in Fig. 4 and by about 38 dB in Fig. 5. From Figs. 4 and 5, it is clear that the agreement between the measured and theoretical values is much better for the first-order gradient estimate than for the second-order gradient estimate. We believe that to achieve any further improvements to the second-order gradient estimate, we have to compensate for the

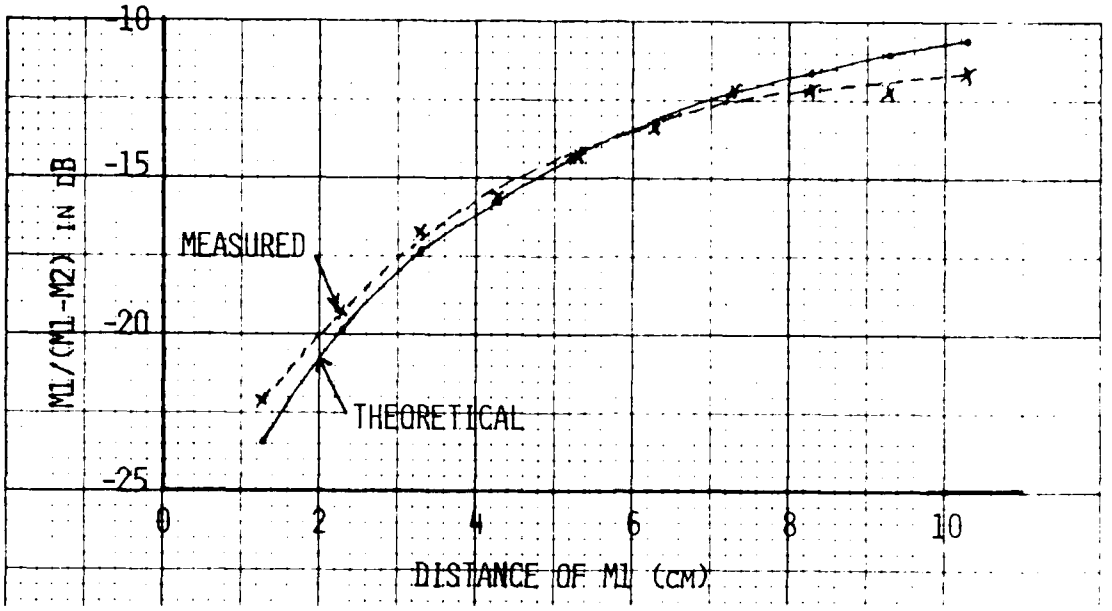


FIG. 4. Plots of theoretical and measured values of M1/(M1-M2) in dB at 465 Hz.

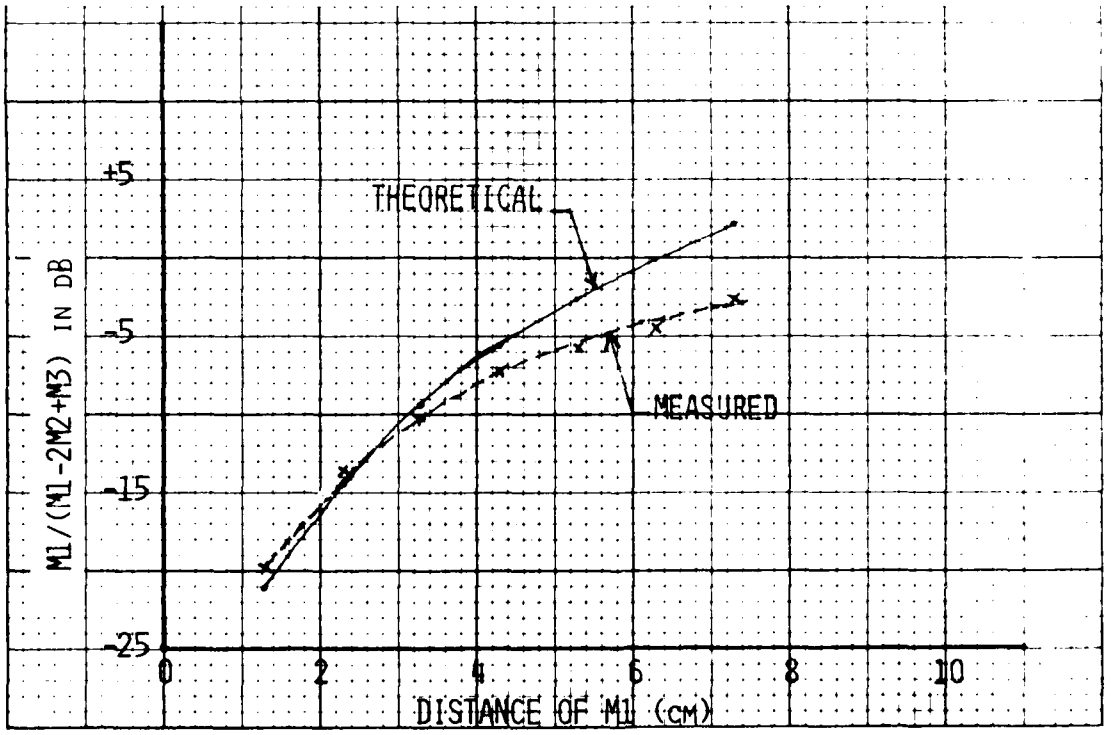


FIG. 5. Plots of theoretical and measured values of M1/(M1-2M2+M3) in dB at 465 Hz.

phase differences between M1-M2 and M3-M2 before they are combined to produce M1-2M2+M3. This involves adding another RC network to the conditioner box and reoptimizing the second-order pressure gradient. This refinement, however, was not performed.

As we were making the measurements described above, we received from the Department of Defense two prototype microphones: ElectroVoice's first-order gradient microphone, EV 985 and Vought's second-order gradient microphone [1]. We repeated the sound-field measurements with the simple source, for each of these two microphones. We then compared the measurements from EV 985 with those from M1-M2 and the measurements from the Vought microphone with those from (M1-2M2+M3). The first-order gradient estimate from our microphone array was uniformly better than the output from EV 985. Our second-order gradient estimate and the Vought microphone output were roughly similar, with Vought being better at 790 Hz and our estimate being better at 1610 Hz. Fig. 6 shows the comparison for the first-order gradient at 790 Hz, and Fig. 7 shows the comparison for the second-order gradient at 465 Hz.

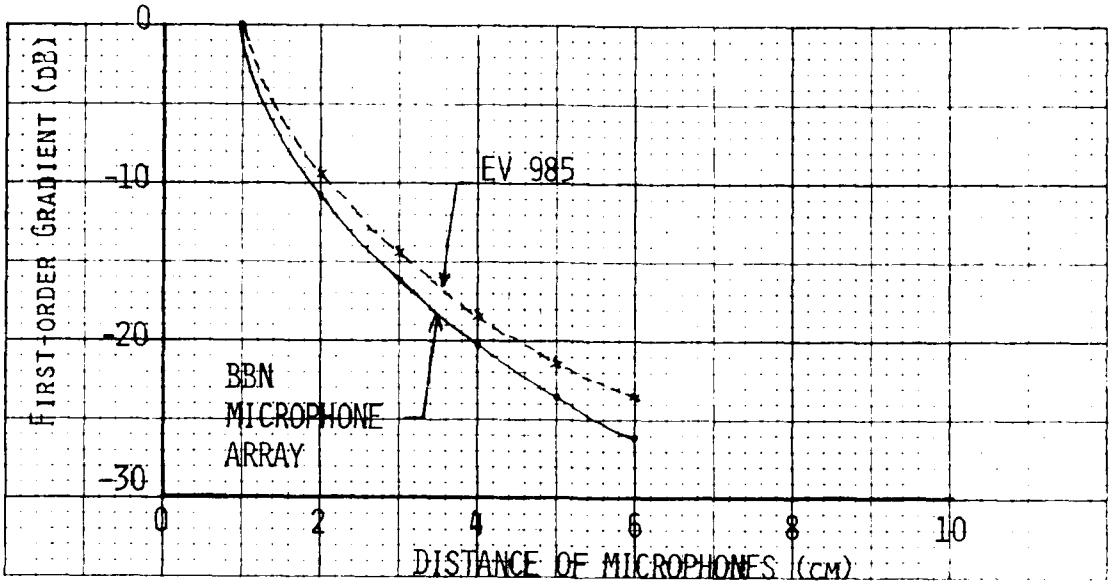


FIG. 6. Comparison of first-order pressure gradients (in dB relative to 1 cm value) produced by EV 985 and BBN microphone array, at 790 Hz.

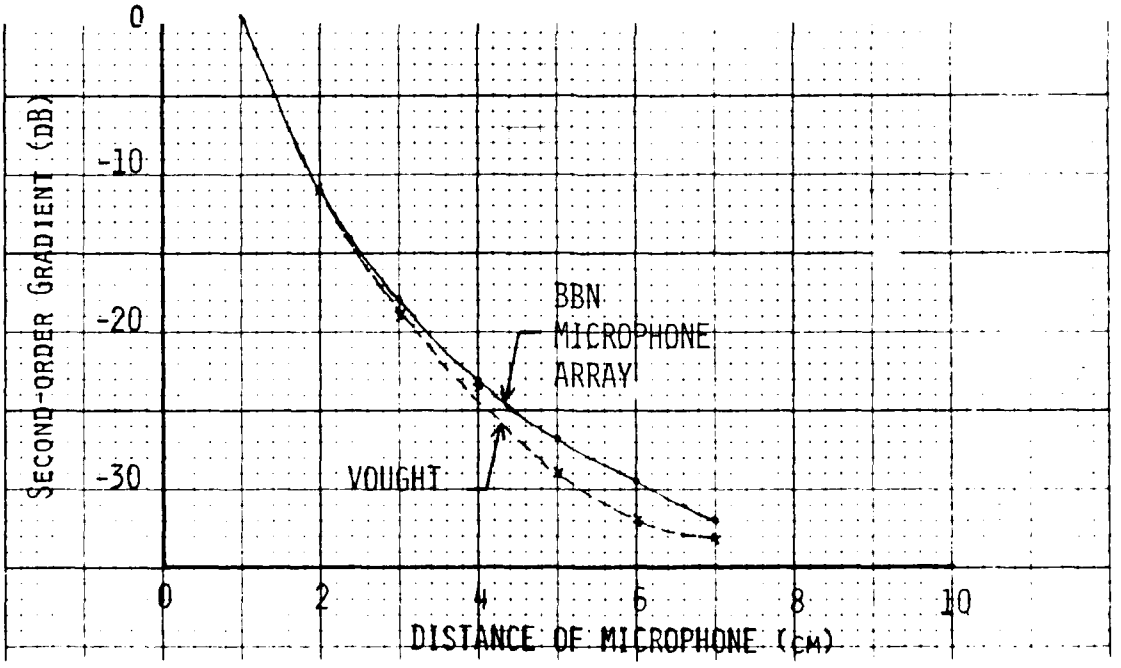


FIG. 7. Comparison of second-order pressure gradients (in dB relative to 1 cm value) produced by Vought microphone and BBN microphone array, at 465 Hz.

3. SOUND-FIELD MEASUREMENTS DURING SPEECH

The purpose of making sound-field measurements during speech is, as we mentioned in Section 1.1, to map out quantitatively the properties of the sound field for a number of different classes of speech sounds in the vicinity of the talker and hence provide a rational basis for the design of multisensor configurations. In this section, we review briefly the prior work in this area; discuss our selection of speech materials, sensors, and sensor positions to be used in the measurements; and describe the experimental setup and the sound-field measurements.

3.1 Review of Prior Work

A review of the literature shows that the sound field in the vicinity of the mouth opening during speech production has not been studied in detail. Some measurements of the sound pressure on the mouth axis at various distances greater than 5 cm have been made by Flanagan [2]. Flanagan also measured the sound pressure at various angles from the mouth axis, but these measurements were at a distance of 30 cm, which is not in the near field except for frequencies below about 200 Hz. At distances closer to the mouth (in the range 1 to 5 cm, say), there are, to our knowledge, no detailed data on the distribution of sound pressure and pressure gradient at various locations around the mouth.

There have been some measurements of the vibration amplitude and waveform at various points on the surfaces of the head and neck during speech production, but these have not been very extensive [3,4]. For example, many years ago Bekesy measured vibration amplitudes over the neck and face surfaces, and displayed his results as contour plots [3]. As might be expected, amplitudes were greatest on neck surfaces near the larynx. Accelerometers have also been used to pick up signals from the sternal notch [5], from the forehead [4], from (or near) the wall of the ear canal [6,7], and from the teeth [8]. However, there has, to our knowledge, been no detailed study of the surface amplitude or velocity or acceleration, and of the spectrum of this signal, for different speech sounds.

3.2 Selection of Speech Materials

The utterances to be produced by each speaker during sound-field measurements should contain an adequate sample of individual speech sounds that are likely to show special characteristics at different transducer locations, relative to the characteristics they exhibit for a reference microphone some distance from the lips. Furthermore, the speech sample should have long-term characteristics that are representative of the average characteristics of conversational speech.

The speech sounds that are likely to show special characteristics when transduced by a close-talking microphone

include sounds for which the location and characteristics of the source of sound radiation deviates significantly from the source at the (unprotruded) lips. These include sounds produced with substantial lip rounding or lip spreading (such as [u, ɔ, i]), nasal sounds for which there is substantial radiation from the nose and possibly from the neck surface, voiced obstruents (such as [b,d,g,z]) for which there may be appreciable low-frequency radiation from the neck surface, and fricative consonants and noise bursts that contain high-frequency energy with a directional characteristic for the radiation from the mouth. The list of utterances we selected is given below.

Isolated Vowels: [i] [æ] [a] [u]

These vowels represent different sizes of mouth opening and different degrees of lip rounding and spreading.

Brief Utterances:

(a) a bat
    a dad
    a gap

The initial consonants (obstruents) in these words will usually be voiced, and thus there will be sound radiation from the neck surface.

(b) a mat
    a knot

These words illustrate nasal consonants in word-initial position.

(c) a man
a mean
can't
hint

These words illustrate nasal consonants in syllable-final position, as well as the vowel nasalization that occurs in the vowels preceding these consonants.

(d) a seat a pot
a sheet a tot
a feat a cot
a church

These words contain examples of voiceless fricative consonants and voiceless stop consonants.

(e) a zoo
a vote
a judge

Voiced fricative consonants occur in these words.

Sentences:

Print a draft of this letter.

I think someone's made a mistake.

Curt will put the box on my desk.

There's room for five more lines in the footnote.

Judy will lose the game of chess.

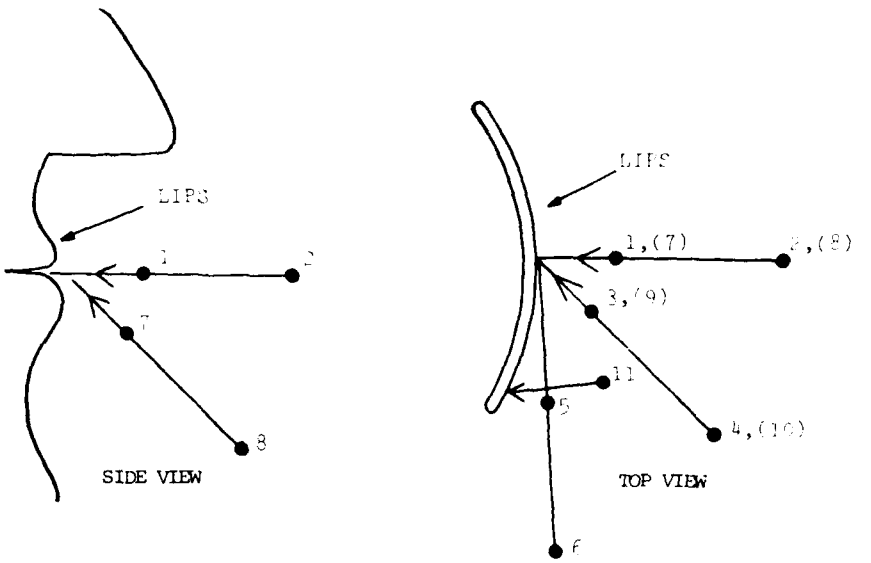
We need to make a bowl of soup.

We selected these six sentences since they contain a number of examples of each of the consonant classes represented in items (a) through (e) above, and since they are representative of running speech. Thus, the sentences can be used to examine the properties of individual speech sounds as well as the average properties over continuous speech.

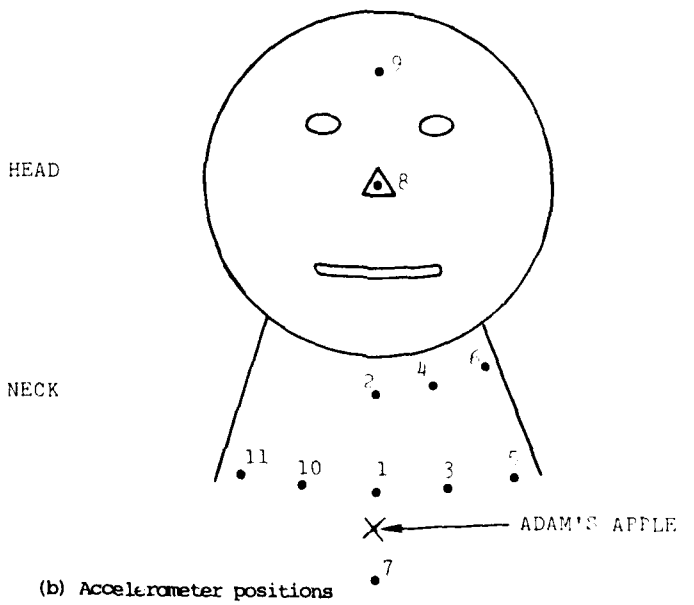
3.3 Sensors and Sensor Positions

The sensors we chose to use for the sound-field measurements are: a reference B & K condenser microphone, the three-microphone array, the first-order gradient microphone EV 985, the Vought second-order gradient microphone, and a BBN 501 accelerometer. As we described in Section 2, the three-microphone array measures sound pressure (M1), first-order pressure gradient (M1-M2), and second-order pressure gradient (M1-2M2+M3).

We decided to locate the reference pressure microphone directly in front of the subject and in line with the center of the subject's mouth, at a distance of 30 cm. We chose 11 positions for locating the microphone array or EV 985 or Vought microphone. These positions are shown in Fig. 8(a). The near positions are at a distance of about 1 cm, and the far positions are at a distance of about 3 cm. (A point at a distance of 1 cm from a simple source is in the near field for frequencies below 5500 Hz, whereas a point at a distance of 3 cm is in the near field below about 1800 Hz.) The positions 1 and 7 (or 2 and 8 or



(a) Microphone positions



(b) Accelerometer positions

FIG. 8. Sensor positions used in our sound-field measurements during speech.

3 and 9 or 4 and 10) are different in the following way. The microphone axis is horizontal for position 1 and is inclined upward at an angle of 45 degrees for position 7. (The microphone axis is along the length of the microphone array; it is the line perpendicular to the front port and passing through the center of the front port, for EV 985; it is the line passing through the center of the front and the back ports, for the Vought microphone.) For positions 1-10, the microphone axis points to the center of the mouth opening. For position 11, the microphone axis points to the corner of the mouth. The 11 positions we chose for the accelerometer are shown in Fig. 8(b). Position 8 is the side of the nose, and position 9 is the center of the forehead. Positions 1 and 3 are above the Adam's apple, and position 7 is below it.

3.4 Sound-Field Measurements

We made the sound-field measurements inside the anechoic chamber. The subject wearing the helmet was seated on a chair in the chamber. All the measuring apparatus was set up on a table outside the chamber. As mentioned above, we decided to record signals from seven sensors: reference B & K condenser microphone, three sensors (M1, M1-M2, and M1-2M2+M3) of the microphone array, BBN 501 accelerometer, EV 985, and Vought microphone. However, we decided against the simultaneous tape recording of the seven sensor signals as it would not be possible to locate in front of

the mouth, without any interference, the microphone array, EV 985, and the Vought microphone (each one having its windscreen). Also, with the helmet-mounted boom we have, it is convenient to use only one of these three microphones at a time. So we decided to perform the tape recordings in three sessions: 1) microphone array, accelerometer, and reference microphone (5-channel tape recording); 2) Vought microphone and reference microphone (2-channel recording); and 3) EV 985 and reference microphone (2-channel recording). The signal from the reference microphone, located always at the same distance of 30 cm from the subject's mouth, will serve as the anchor for the across-session comparisons.

For the first session, we used a 7-channel FM tape recorder Racal Store 7-DS for the simultaneous recording of the 5 sensor signals. We operated the tape recorder in its wideband mode and at a tape speed of 7 1/2 ips. With this choice, the bandwidth of the recorder is 0-5 kHz and the signal-to-noise ratio is 48 dB. For the second and third sessions, we used a Kudelski Nagra IV-SJ stereo tape recorder for recording the 2 sensor signals. Each sensor signal was connected to an Ithaco (calibrated) amplifier (Model No. 453M103), which was in turn connected, through a switch-box, to the appropriate input channel of the tape recorder. The switch-box was used to select the sensor signals or a multichannel marker pulse. The signal being taped was monitored one channel at a time using an oscilloscope and using a

headset. For each of the eleven sensor positions, the subject was asked to read the list of speech utterances, and the amplifier gains were chosen to fill the channel input ranges without overloading.

As speech materials to be recorded, we used the list of isolated vowels, words, and sentences given above in Section 3.2 and included here as Table 1. We divided this list into 5 sections A-E, as shown in Table 1. The subject was asked to start each section by reading the section name (e.g., Section A). Just prior to recording each section, we recorded a multichannel marker pulse (produced by switching in and out a fixed DC voltage); this pulse can be used for the purpose of synchronizing the various channel signals (after digitization) with each other. (Such synchronization is not necessary if we digitize using a multichannel A/D converter. See Section 4.2.)

To check out our measurement procedure described above, we carried out a "dry-run" for Session 1, using the microphone array and the accelerometer. We found out that without a windscreen (or puffscreen) for the microphone array, we needed to use sufficiently low gain settings to prevent any overloading of the tape recorder channels. But, the resulting tape recording was of poor quality with a low signal-to-noise ratio. Higher gain settings avoided this problem at the expense of moderate-to-severe overloading, mostly caused by breath noise. Guided by this experience, we decided to use windscreens for the

(A)	
	i (Beat)
	æ (Bat)
	a (Bob)
	u (Boot)
	a bat
	a dad
	a gap
(B)	
	a mat
	a knot
	a man
	a mean
	can't
	hint
(C)	
	a seat
	a pot
	a sheet
	a tot
	a feat
	a cot
	a church
(D)	
	a zoo
	a vote
	a judge
	Print a draft of this letter.
	I think someone's made a mistake.
	Curt will put the box on my desk.
(E)	There's room for five more lines in the footnote.
	Judy will lose the game of chess.
	We need to make a bowl of soup.

TABLE 1. A list of speech utterances used in sound-field measurements.

microphones to reduce the effect of breath noise. We made windscreens from open-cell foam for the microphone array and for EV 985. (The Vought microphone came with its own windscreen.)

For Session 1 involving the microphone array and the accelerometer, we recorded five subjects, 3 males and 2 females. For Sessions 2 and 3, involving, respectively, EV 985 and Vought, we recorded only two of the five subjects, 1 male and 1 female. For each session, we located the sensors at each of the eleven positions (depicted in Fig. 8); for each position, the subject read all 5 sections of the list given in Table 1. As we mentioned above, the subject wore the helmet-mounted boom, and the microphone was affixed to the boom with clamps. The accelerometer was attached to the throat with an easily removable double-stick tape. For the near positions (1, 3, 5, 7, 9, and 11; see Fig. 8 (a)), we located the microphone as close to the lips as possible without the windscreen touching the lips even for sounds that require protruded lips (e.g., [u] as in zoo).

We measured the microphone distance as the distance to the microphone (without the windscreen) from the center of the vertical line passing through the centers of the lower and upper lips. After recording a near position, we moved the microphone away by about 2 to 2.5 cm to obtain the corresponding far position. After each position was completed, we took polaroid pictures of the subject with the sensors clearly showing and with a tape measure placed next to the subject. These pictures should

be useful for checking the location and orientation of the sensors. The measured distances for one male subject are given in Table 2 for all three recording sessions. (We note that our subsequent amplitude and spectral investigations, described in Section 4, were performed on the measured data of this subject only.)

Position No.	BBN Array	EV 985	Vought
1	0.7	2.2	2.4
2	3.0	4.6	4.8
3	0.6	2.4	2.2
4	3.0	5.3	4.3
5	2.6	4.0	4.4
6	4.8	5.7	6.3
7	0.5	2.3	2.0
8	3.0	4.3	4.8
9	0.6	2.3	2.5
10	3.0	4.4	4.5
11	2.4	2.6	3.6

TABLE 2. Distances (in cm) of the three microphones to the center of the mouth for each of the 11 positions, all measured during sound-field measurements for one male subject.

Finally, we recorded calibration tones as follows. For each of the microphones (reference, three-microphones array, EV 985, and Vought), we used our simple source at 465 Hz (see Section 2.8), located the microphone (other than the reference) at a distance of 2 cm from the source, and set the level so the reference microphone at 30 cm from the source sensed a sound pressure of 74 dB. The calibration tone we recorded is the output of the respective sensor in this setup. The calibration

tone for the accelerometer was obtained from a vibration calibrator that was set for an acceleration of 1 g at 100 Hz.

4. ANALYSIS OF THE MEASURED SPEECH DATA

In this section, we describe the methods we used for analyzing the tape-recorded speech data and present the results from these analyses. Initially, we studied the spectrograms of a few sentences from one subject and for the different sensors, and the results of this study were used in making plans for subsequent computer analyses of the data (Section 4.1). After digitizing the measured speech data of one subject using our multichannel A/D facility (Section 4.2), we performed informal listening tests (Section 4.3), long-term spectral analysis (Section 4.5), and short-term spectral analysis (Section 4.6) on selected subsets of the digitized data. To evaluate theoretically the speech intelligibility performance of the different sensors under aircraft cockpit noise, we first estimated the sensor responses under simulated cockpit noise (Section 4.7) and computed the articulation index scores of the sensors (Section 4.8). In these spectral and articulation index analyses, we compared the different sensors for a given position as well as the different positions for a given sensor.

4.1 Analysis Using Spectrograms

We made spectrograms of a few sentences for each of the 7 sensor signals we recorded for one position (Position 1 in Fig. 8). From a comparative study of these spectrograms, we made several observations. First, the spectrograms for M1 and the

reference microphone are quite similar. Second, the spectrograms for M1-2M2+M3 show the presence of low-level broadband noise, but they are otherwise similar to the spectrograms for M1. This indicates that the second-order pressure gradient estimate obtained from the three-microphone array is not good. Therefore, we decided not to include the data from M1-2M2+M3 in all our subsequent investigations. Henceforth, when we refer to second-order gradient microphone, we shall mean the Vought microphone. Third, as gradient microphones are directional, we expect that their outputs will depend strongly on the location of the sound source. As we expected, nasals, nasalized vowels, and prevoicing in voiced obstruents are much weaker for M1-M2, EV 985, and Vought than for M1 and the reference microphone. Fig. 9 shows the spectrograms of part of the sentence "We need to make a bowl of soup", for M1, M1-M2, and Vought, for position 1. Although the reproduced spectrograms in Fig. 9 do not show the amplitude relations very well, it is clear that the nasal murmur for [n] in "need" and [m] in "make" and the voice bar preceding the release for [b] in "bowl" are much weaker in relation to the adjacent vowel for the gradient microphones than for the pressure microphone.

Fourth, as Fig. 10 illustrates, the accelerometer signal contains substantial energy up to about 2.5 kHz and little energy above this frequency. The accelerometer picks up the first and second formants and voice bars, but does not sense the

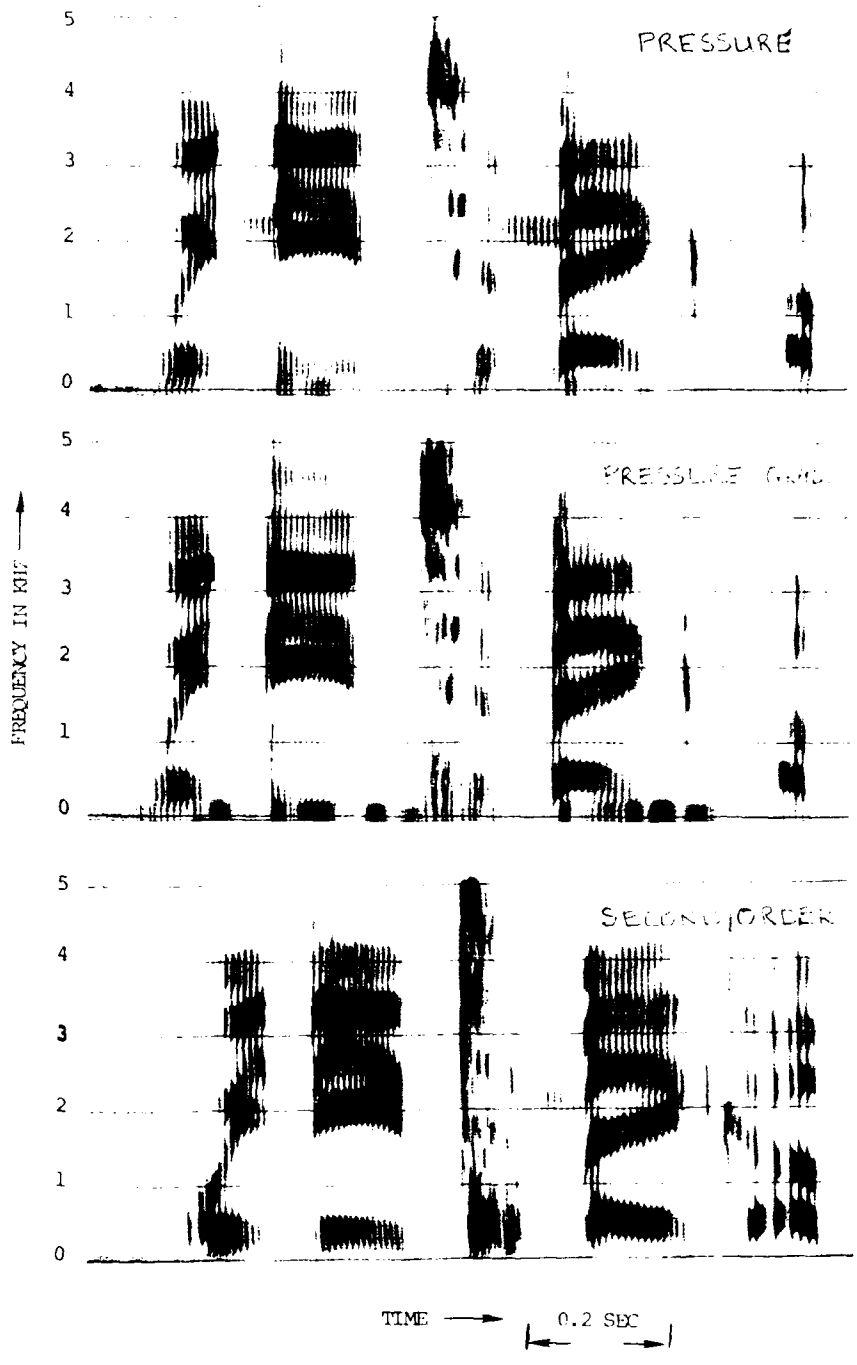


FIG. 9. Spectrograms of the signals from pressure and pressure gradient microphones, for position 1 and part of the sentence "We need to make a bowl of soup."

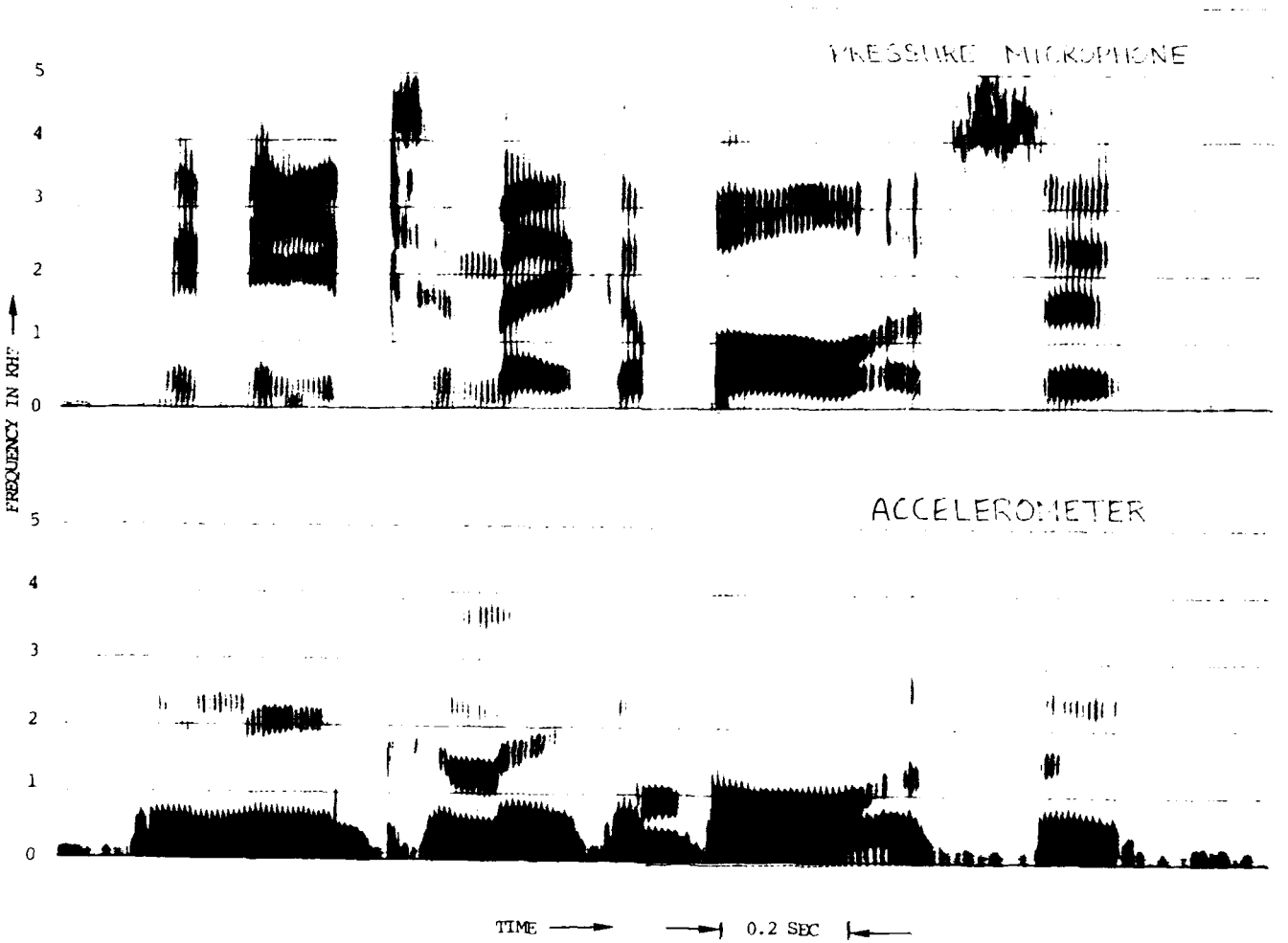


FIG. 10. Spectrograms of the signals from the pressure microphone and the accelerometer, for position 10 and the sentence "We need to make a bowl of soup."

high-frequency information of the fricatives ([s] in "soup" in Fig. 10) and bursts ([t] burst in "to" in Fig. 10).

From the results of our spectrographic analysis, we made a list of items to work on, as part of our computer analysis of the tape-recordings. We planned to examine the spectra and relative amplitudes at selected points within a number of utterances, for the items listed below:

- Four isolated vowels [i æ a u] recorded from the accelerometer at 11 different locations on the head and neck.
- Voice bar in a voiced stop and nasal murmur during a nasal consonant, recorded from selected accelerometer locations.
- Selected fricatives and stop bursts, recorded from selected accelerometer locations.
- Several vowels recorded from each location of the three types of microphones (pressure, pressure gradient, second-order gradient).
- Nasal murmurs, voice bars, fricative consonants, and noise bursts for selected microphone locations. (We must examine their spectra and amplitudes in relation to adjacent vowel.)

Also, we planned to compute the long-term average spectra over several sentences, for all sensors and locations.

4.2 Data Digitization

We digitized all the recorded data for one male subject, using our multichannel A/D facility, at a sampling frequency of 10 kHz per channel. With the multichannel A/D, the digitized

individual channel signals are synchronized in time. The digitized data was written out as interleaved waveform files in which, for each sampling period, the samples of the different channels are stored adjacent to each other. The number of channels digitized is stored in the file header.

The analog equipment used for the A/D process was: a Racal Store 7-DS FM recorder (for 5-channel recordings), a Nagra stereo recorder (for 2-channel recordings), Ithaco amplifiers (one per channel), and Krohn-Hite variable-cutoff lowpass filters (one per channel, with cutoff set to 4.7 kHz to avoid aliasing).

Because the recording levels varied from one sensor position to another, it was necessary to adjust the gain on the Ithaco amplifiers to achieve a good dynamic range without clipping the digitized data. However, within each position, gain settings remained constant for all 5 sections of utterances. These settings were recorded for future reference.

We then divided each interleaved waveform file into single-channel files for compatibility with our D/A and display software. This was accomplished with a program that divided the N-channel interleaved waveforms into N individual buffers and wrote them to files that were mnemonically named to encode the microphone position, utterance section, and microphone type. Also, the composite recording and digitizing gains (in dB) were output to the individual file headers, to facilitate access to this information by our display and analysis programs.

4.3 Informal Listening Tests

Using our flexible D/A program, we played out the digitized speech files and informally listened to several utterances recorded from different sensors and positions, with the aim of detecting audible variations as a function of a given sensor's location.

The most noticeable effect of position change was heard for the accelerometer. In particular, position 10 (see Fig. 8(b)) located just to the right and slightly above the glottis, produced the most intelligible speech. Position 7, below the glottis, was barely intelligible. The intelligibility of the other locations fell in between these two extremes. Position 8, on the nose, caused significant increases in volume for nasals and nasalized sounds.

The other sensors (pressure and gradient microphones) did not produce so much variation over the range of positions in terms of speech quality. However, one major effect was the change in low frequency noise as the sensor was moved in or out of the breath stream. Plosives and fricatives producing a significant breath stream were audibly distorted for sensor positions directly in front of the subject's mouth. Sensor locations to the side of the mouth did not have this problem. We found that highpass filtering at about 200 Hz substantially reduced the perceived level of the breath stream noise.

We observed another position-related effect for the first order gradient microphone. As the microphone was moved farther from the speech source, nasal sounds were somewhat less natural-sounding with a stop-like quality.

4.4 An Interactive Display Program

We developed an interactive waveform and spectral display program for the analysis and investigation of our digitized database. The program provides simultaneous display of two waveforms and two spectra, corresponding to two sensors or sensor positions. The display program takes advantage of the bit-map capabilities of BBN's BitGraph graphics terminal to manipulate individual display windows independently.

A typical application of the program for this project is comparison of spectra of a given sound recorded with two different sensors at the same position and using the same utterance. In this situation, the user can specify the appropriate files and display the two waveforms. He can control the time scale of the displays and scroll each waveform independently to view different portions of the file or to achieve time alignment of the two displays. To compare the spectra at a certain point in the two utterances, the user moves a cursor to the desired position and types the "spectrum" command. The program computes a DFT on Hamming-windowed frame at the cursor position and displays the spectrum or spectra above

the waveforms. The spectra are superimposed on the same axes for ease of comparison. The program provides the option to display the short-term spectrum of each waveform in any of several modes: unnormalized, normalized with respect to the recording/digitizing gain, or normalized with respect to the reference microphone signal for the same frame of the given utterance. The latter two normalizations may also be combined. The program allows the user to examine spectral amplitudes of the two displayed spectra at any frequency. This is accomplished by moving a cursor to the desired point on the spectral display. The program automatically updates a numeric display of the frequency and amplitudes to their current values.

The program has the capability for computing the long-term average spectrum, over only the speech portions of a given file. Even if a file contains large amounts of silence, these portions can be excised from the computation so they do not corrupt the spectral average. The program also computes a linear prediction fit to the long-term spectrum as a means of smoothing it. The number of poles in the linear prediction model is a parameter that may be specified at run time. Both unsmoothed and smoothed spectra may be displayed in the same manner as described above.

4.5 Long-Term Spectral Analysis

We computed the long-term average spectrum of each sensor signal at each location by averaging the short-term spectra

(computed using 512-point FFT on 256-sample Hamming-windowed speech frames) over about 10 seconds of speech. (We used five of the six sentences given in Table 1.) Before we compared the long-term average spectral amplitudes over the eleven positions of a sensor, it was necessary to verify that the observed differences were in fact a result of position changes, rather than differences in the way the utterances were spoken. We smoothed the long-term average spectrum of the reference microphone signal, computed for each of the eleven repetitions (each corresponding to a different sensor location), using a 14th order linear prediction fit. Visual comparison of plots of these spectra revealed that while the overall level varied as much as 9 dB, the spectral shape remained within 3 dB throughout the 5 kHz frequency range. The agreement was usually closer, within 1 or 2 dB, for frequencies from 1 to 3 kHz. Thus, if we normalize for overall level, it is safe to attribute the results reported below to changes in sensor position and type rather than significant variations in the speaking of the utterances themselves.

We analyzed the average spectral data to determine how the spectra vary as a function of sensor type, location, and orientation, and to examine any frequency dependence that might exist in those variations. From visual inspection of the reference microphone average spectra, we selected six frequencies, which were normally at the center of spectral peaks, to perform the analysis. These peaks also corresponded to

regions of particularly useful information in the spectrum: 120 Hz (average F0), 240 Hz (average second harmonic), 450 Hz (center of the typically largest peak in the average spectrum of speech), 1200 Hz (upper edge of useful information for most accelerometer positions), 2400 Hz (upper edge of most information in average spectrum), and 3200 Hz (contains some information on contributions of fricatives and noise bursts).

Below, we report the results of our long-term spectral analysis, first for pressure and pressure gradient microphones and then for the accelerometer.

4.5.1 Average Spectra for the Pressure and Pressure Gradient Microphones

In Fig. 11, we have plotted the average spectral amplitude of the gradient microphone M1-M2 as a function of the microphone position for three frequencies. Notice that the near positions (1, 3, 5, 7, 9, and 11) are plotted at the left and the far positions (2, 4, 6, 8, and 10) are plotted at the right. As expected, the spectral amplitudes are larger for the near positions than for the far positions. Of the near positions, 11 is particularly bad, and 3, 7, and 9 seem to be good. We obtained similar results for M1, EV 985, and Vought.

We then lumped together the data for the near positions and for the far positions to examine the effect of source-to-sensor distance on sensor response. Position 11 was not included among

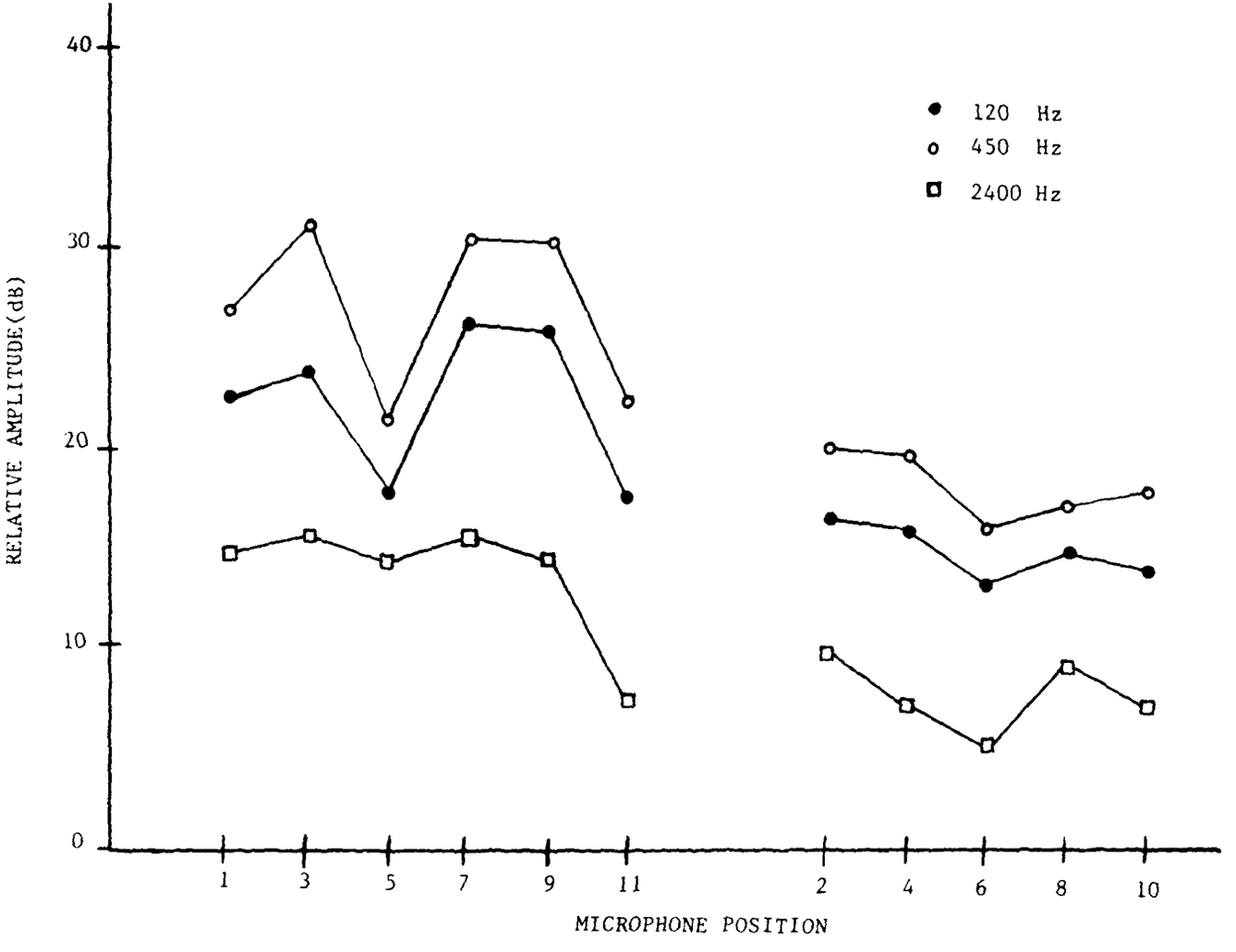


FIG. 11. Long-term average spectral amplitude plotted as a function of position, for M1-M2.

the near group because it was oriented toward the corner, rather than the center, of the mouth. We observe from Table 2 that position 5 had actually been placed slightly farther from the mouth than the other near positions (to avoid contact with the speaker's lips). So we eliminated positions 5 and 6 from the near and far groups, respectively, to remove any bias introduced by this effect. Response at position 5 was in fact substantially weaker than that of the other near positions, as illustrated in Fig. 11; however, we do not know whether this should be attributed to the effect of distance or its location at the side of the mouth.

The results for two frequencies are shown in Fig. 12 for M1, M1-M2, and Vought and in Fig. 13 for M1-M2 and EV 985. The data is in terms of dB relative to the reference microphone spectrum at the same frequency. We can draw several conclusions from the data (including data not shown in Figs. 12 and 13). First, we observe that the response of all four sensors drops as the distance is increased, at all frequencies. We recall that in the near field of an ideal point source, the sensor output drops off proportionally to $1/r$ for the pressure microphone, $1/r^2$ for the first-order gradient microphone, and $1/r^3$ for the second-order gradient microphone, where r is the distance. Although the human vocal mechanism is not a point source, we see that the higher order microphones do in fact exhibit the expected greater sensitivity to distance. Second, the decay with distance is also

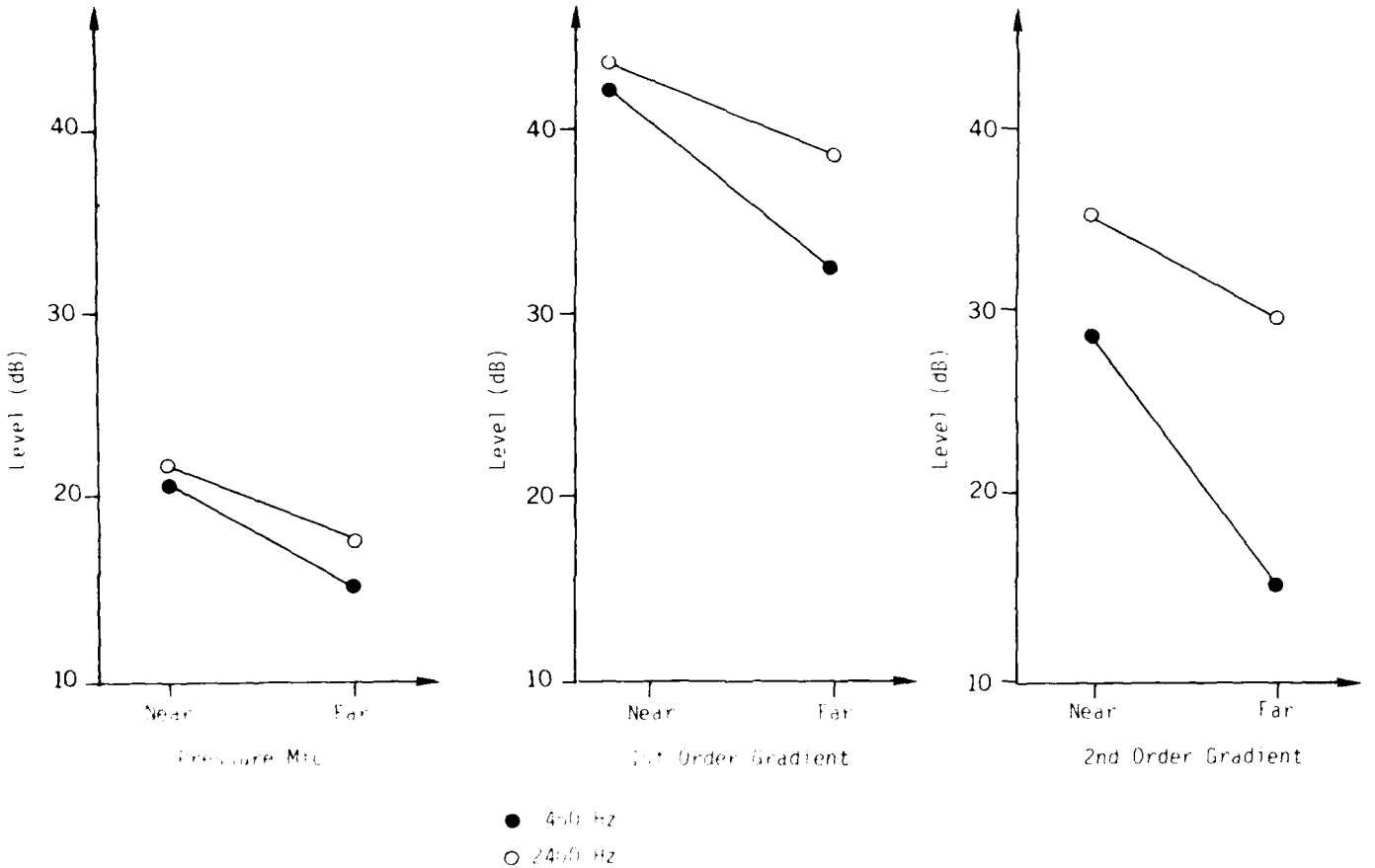


FIG. 12. Effect of source-to-sensor distance, for M1, M1-M2, and Vought.

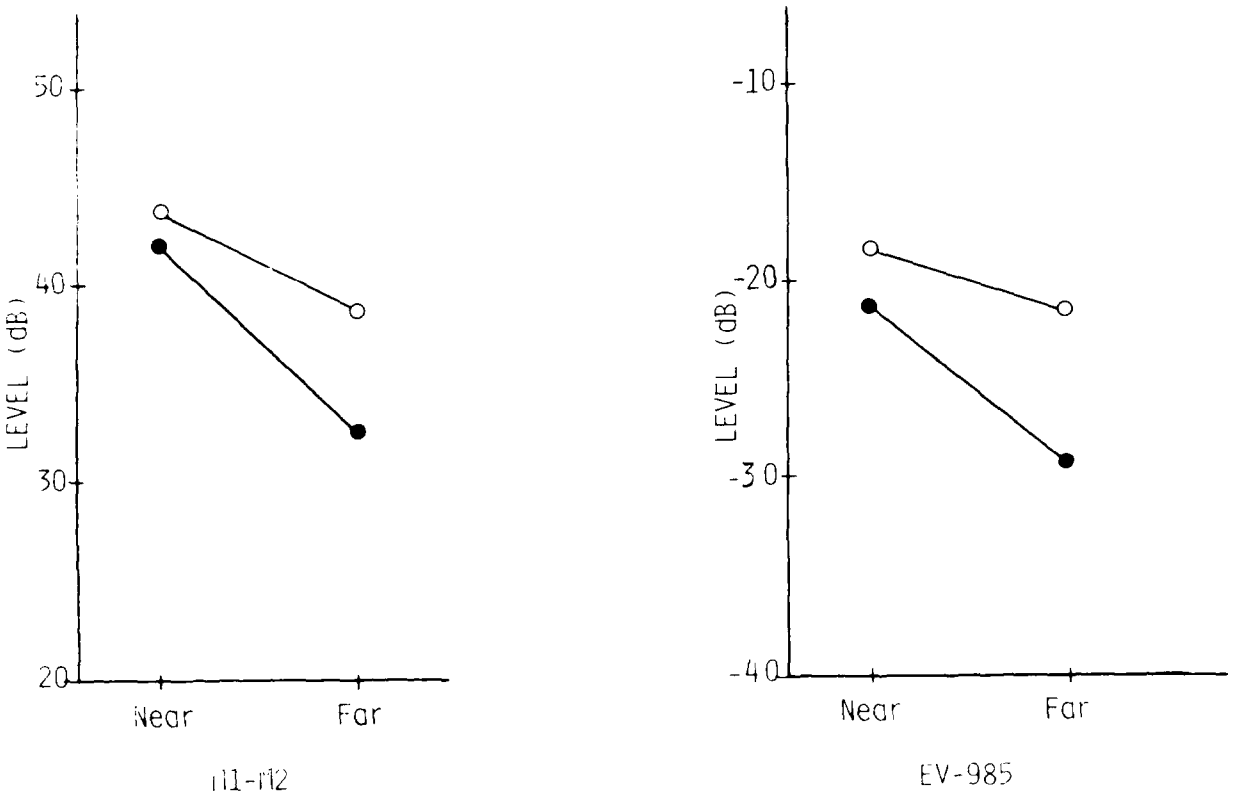


FIG. 13. Effect of source-to-sensor distance, for M1-M2 and EV 985.

frequency dependent. Since both near and far positions are in the near field at lower frequencies, the decay with distance is greater for lower frequencies than for higher frequencies. Clearly, positioning of these microphones is critical; a shift of one or two cm can produce a significant change in level, particularly at lower frequencies. Third, Fig. 13 shows that the decreases in response (going from near to far position) for M1-M2 and EV 985 are, respectively, 5 dB and 3.4 dB at 450 kHz and 9.5 dB and 7.3 dB at 2400 Hz. The smaller decreases for EV 985 also imply that the pressure gradient estimate it provides is less accurate than the estimate from M1-M2.

Next, we examined the effect of sensor orientation. Figs. 14 and 15 show the sensitivity of each of the four sensors to orientation differences. The three near positions (3, 9, and 11) considered are all 45^{o} from the midline (see Fig. 8). We recall the position 3 is horizontal, pointing toward the center of the mouth, position 9 points up at a 45^{o} angle to the center of the mouth, and position 11 is horizontal pointing toward the corner of the mouth. We observe that proper microphone orientation is very important. For all sensors except EV 985, the orientation of position 9 leads, in general, to the highest transduced levels; it is slightly better than the orientation of position 3 and substantially better than the orientation of position 11. For EV 985, however, position 3 gives the best orientation (see Fig. 15).

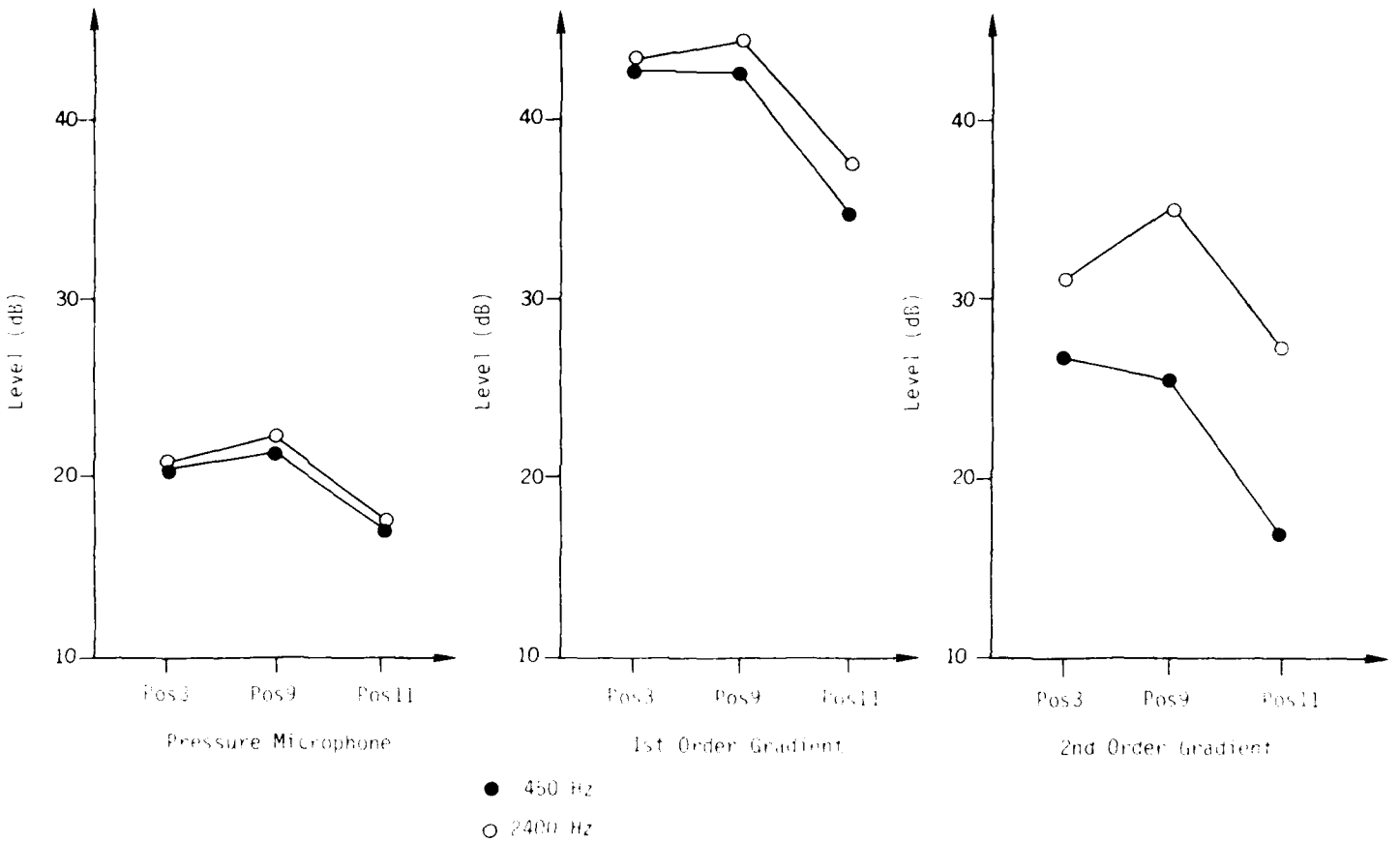


FIG. 14. Effect of sensor orientation for M1, M1-M2, and Vought.

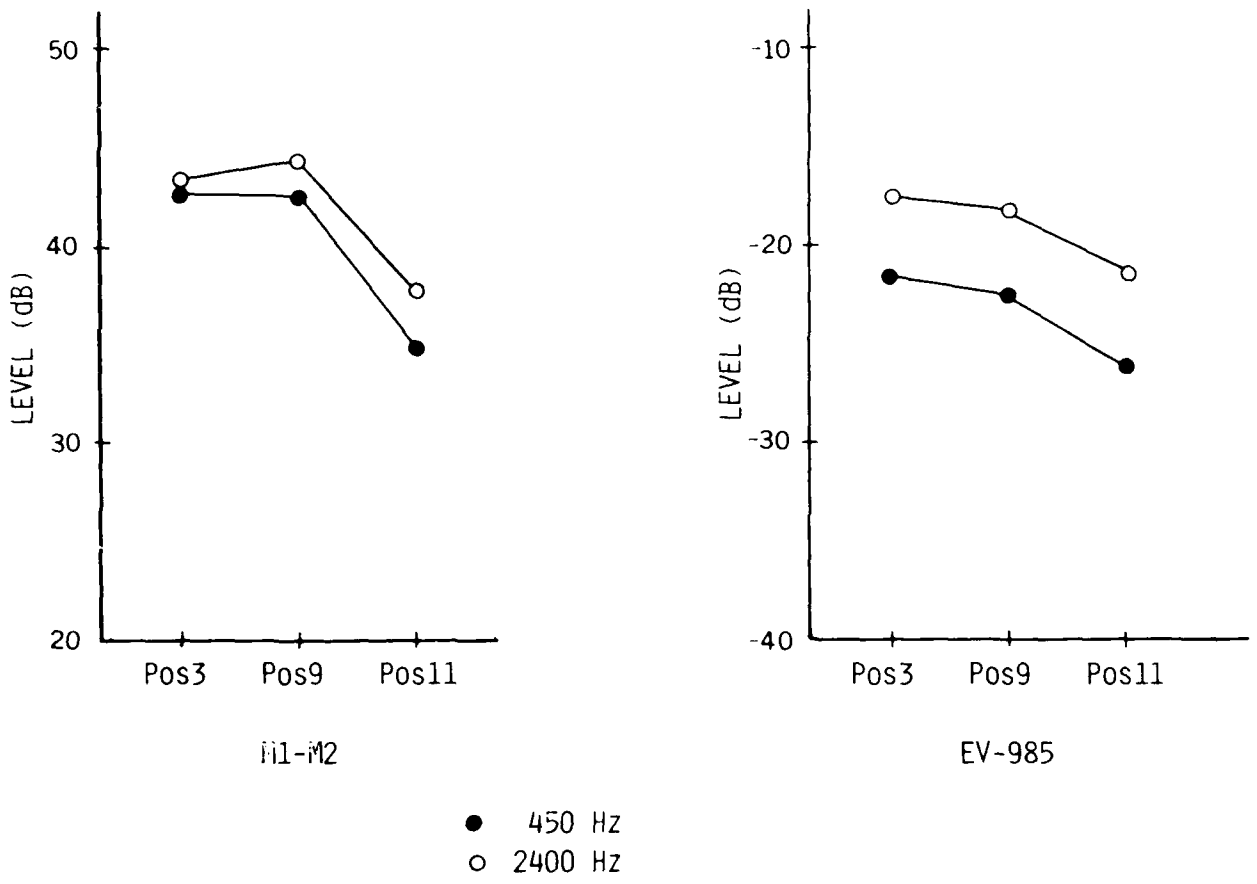
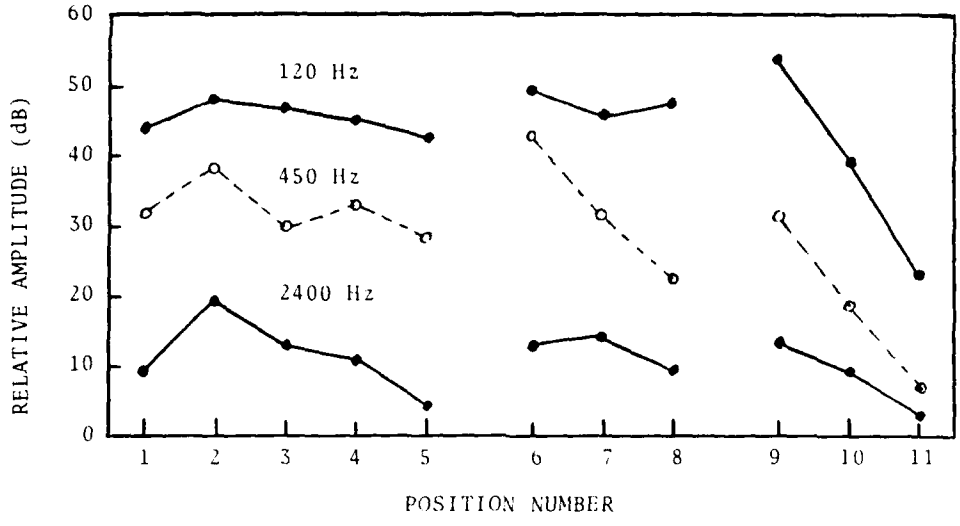


FIG. 15. Effect of sensor orientation for M1-M2 and EV 985.

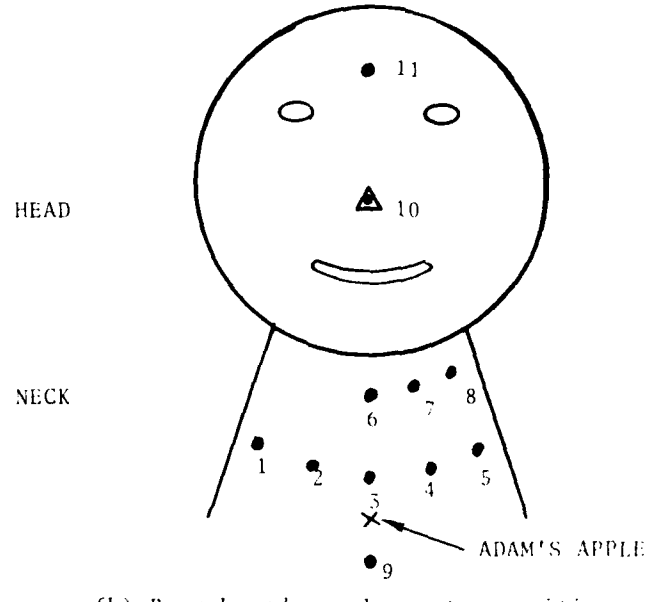
The relatively flat plots for EV 985 in Fig. 15 indicate that this microphone is less sensitive to orientation than is M1-M2. We note that the effect of sensor orientation is also frequency dependent, as evident in Figs. 14 and 15. From Fig. 11, we see that for the far positions, horizontal orientation (positions 2 and 4) is slightly better than 45° orientation (positions 8 and 10).

4.5.2 Average Spectra for the accelerometer

The data on the long-term spectral amplitude for the 11 accelerometer positions at three frequencies are shown in Fig. 16. We have renumbered the accelerometer positions, as shown in the figure, to demonstrate a more logical grouping by position. As expected, the maximum amplitude for the fundamental component (at 120 Hz) is obtained above the sternal notch and below the thyroid cartilage -- position 9 in the figure. There is evidence for some higher frequency speech energy for positions above the larynx, especially for a few to either side of the midline. On the basis of our long-term average data, position 2 (formerly position 10) shows the most high frequency energy. Informal listening tests, as reported in Section 4.3, indicated that this signal was in fact the most intelligible of all accelerometer locations. We do not know whether asymmetry (positions 1-5) with respect to the midline is a speaker-dependent artifact or is an universal property.



(a) Plot of the long-term spectral amplitude.



(b) Renumbered accelerometer positions.

FIG. 16. Long-term spectral amplitude versus position number, for the accelerometer signal.

Note that for all positions, even the best one, amplitude drops substantially for higher frequencies. For the best position, there is little energy above 2000 Hz; for others, the useful band extends only to 1200 Hz or lower. Of course, the accelerometer signal is heavily weighted toward the low frequencies because of the frequency dependent characteristics of sound transmission through the skin. This property has different implications for intelligibility in voice communication systems than for feature extraction. A position that is best suited for pitch extraction (position 9) may not be very intelligible (listening tests showed that it was not), and vice versa. If in a multisensor configuration we decide to use the accelerometer signal for low frequencies (e.g., 0-1 kHz), we see from Fig. 16 that position 6 is the best position.

4.6 Short-Term Spectral Analysis

Analysis of short-term spectra is important for comparing the effectiveness of different sensors and positions in transducing specific phonemes. Our spectrographic analysis, reported above in Section 4.1, has already shown that nasal murmur and voice bar are weaker for the gradient microphones than for the pressure microphone.

For various classes of phonemes, we collected data of short-term spectral amplitudes using our interactive display program, as follows. We located a frame in the steady-state

portion of the desired phoneme and computed the power spectra of the signals from the sensor under study and the reference microphone. We then located and recorded the frequency and amplitude of relevant spectral features for that phoneme. The features we chose to examine are: the fundamental frequency (1st harmonic or FØ) and the first four formants (F1, F2, F3, and F4), for vowels; FØ and 2 FØ (which represent the nasal murmur), for nasals; FØ and 2 FØ (which represent also the voice bar), for other voiced consonants; and the most prominent peak(s) in the upper frequency region where one would expect the highest concentration of energy, for fricatives (e.g., 4 kHz for the phoneme [s]).

In comparing the spectral amplitudes of a given phoneme for different sensor positions, it is important to normalize for the speaking level of the subject. This can be done by computing the spectral amplitude relative to the reference microphone. When the comparisons involve different sensors, it is important to normalize for sensor gains. An effective way of achieving this normalization is to compute the spectral amplitudes of a given phoneme relative to an adjacent vowel; this method normalizes for the speaking level as well.

Our analysis of the spectral data that we collected for the gradient microphone M1-M2 showed that nasal murmur and voice bar were about 8-10 dB weaker (in relation to an adjacent vowel) for M1-M2 than for the reference microphone. The observed difference

was even greater for the Vought second-order gradient microphone. A complete quantitative analysis of the short-term spectral data was not performed in this contract because of limited funds.

Considering the accelerometer, we know that the transduced signal lacks high-frequency information. Fig. 17 shows the short-term DFT spectrum of the vowel /i/ in the word "need" from the sentence "We need to make a bowl of soup," for the accelerometer signal at position 2 (see Fig. 16) and for the reference microphone. The enhanced fundamental is evident for the accelerometer signal, as is the second formant peak, but there is little speech energy above the "noise" at higher frequencies. A spectral zero is apparent between F1 and F2 (at about 1300 Hz), as would be expected for a vowel sensed just above the glottis. For other positions, we observed the loss of information above 1000-1200 Hz, eliminating F2 altogether. Also, we found that the locations on the nose and the forehead sensed the nasal murmur much better than did any of the other locations, but these two locations, however, produced minimal information for non-nasals.

4.7 Theoretical Computation of Sensor Responses in Cockpit Noise

Spectral analyses reported above did not involve any acoustic background noise since the speech data they used were measured in a noise-free anechoic chamber. For practical applications, however, it is necessary to compare the performance

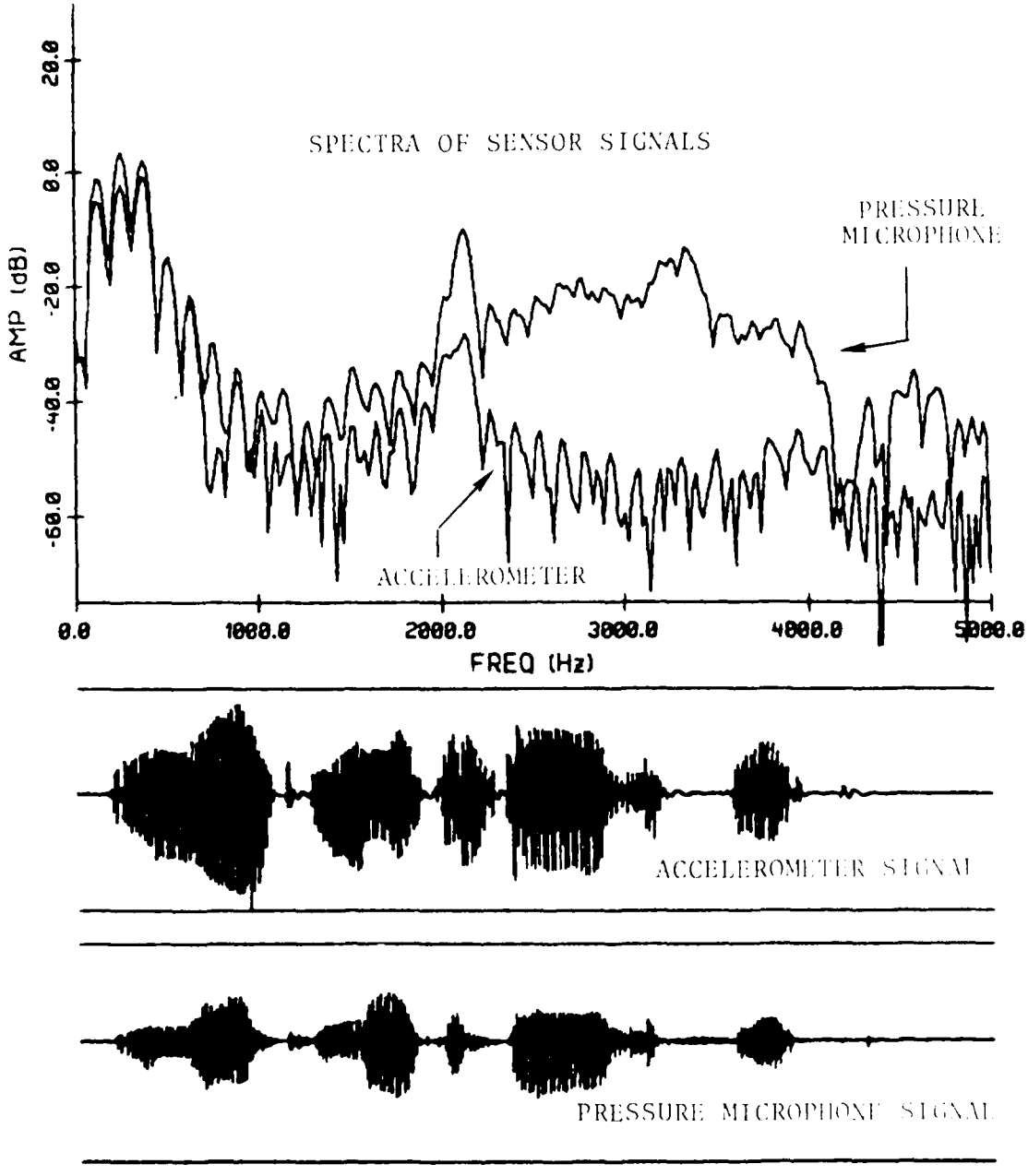


FIG. 17. Waveform and short-term spectral comparisons of the signals from the accelerometer and the reference pressure microphone.

of various sensors in a noisy environment. We decided to consider the cockpit noise of a typical fighter aircraft in view of its relevance to the sponsor's application. To evaluate sensor performance, we decided to use the articulation index (AI) [9] as the criterion. The AI is a theoretical measure of speech intelligibility, and its computation (see Section 4.8) requires the long-term average speech spectrum and noise spectrum. In this section, we describe how we theoretically computed the long-term average spectrum of the noise transduced by each of the different sensors (except the accelerometer, which is essentially insensitive to ambient noise).

To determine a representative level and spectrum of noise encountered in the cockpit of typical fighter aircraft, we examined about 20-30 different noise spectra measured in the cockpit of an F-15 aircraft under various operating conditions, including takeoff, climb, level flight, and maneuvering flight. (These measurements were made as part of a previous BBN contract for Lincoln Laboratory [10].) We obtained averages and standard deviations of the levels in a number of one-third octave bands, over these various measurements. The results of this process are shown in Fig. 18. The vertical bars are separated by two standard deviations. We computed the overall level of the cockpit noise spectrum in Fig. 18 to be about 107 dB (re 0.0002 microbar).

We then recorded on audio tape measurements of the sensor

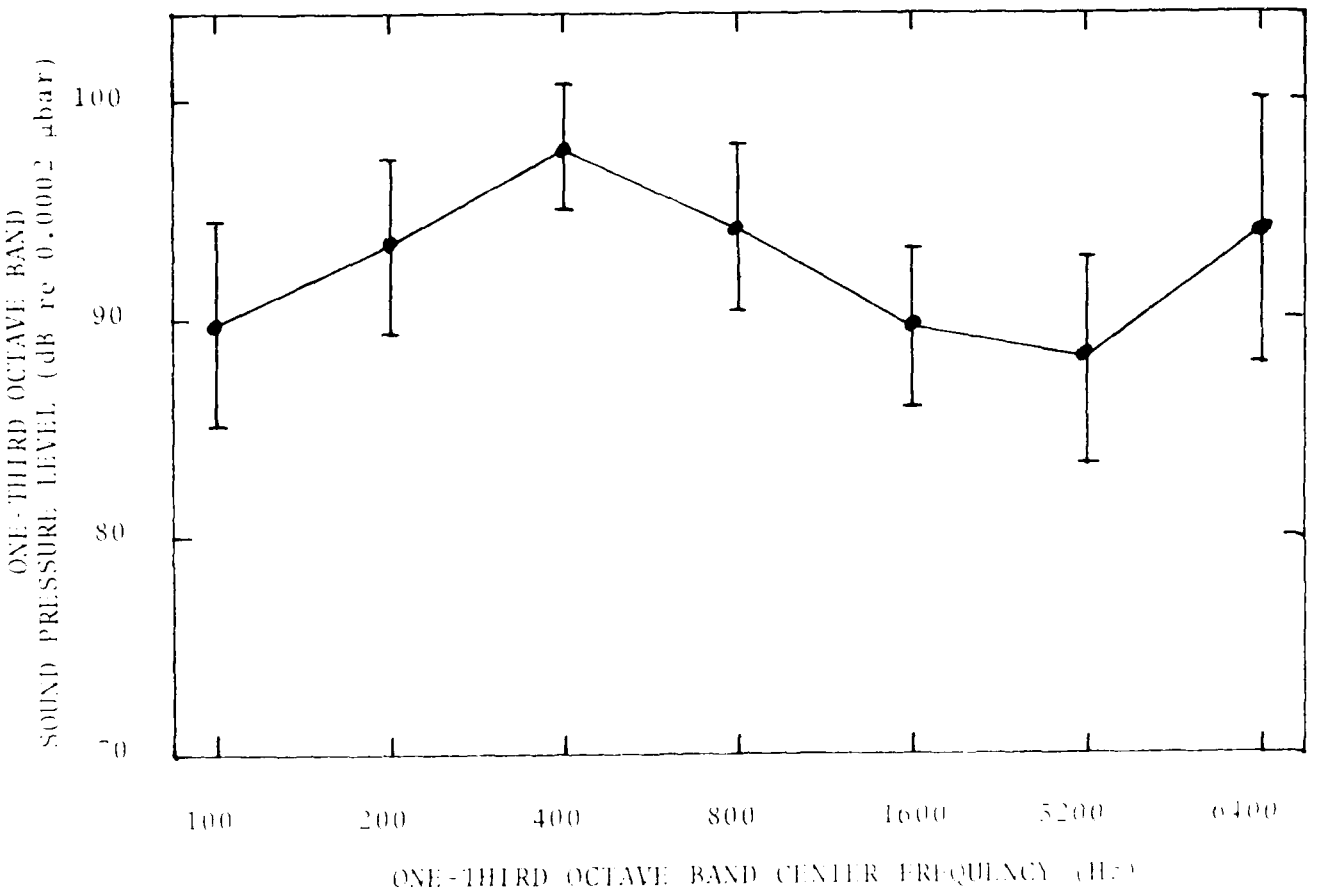


FIG. 18. Average and standard deviation of a number of measurements of sound pressure levels in one-third octave bands in the cockpit of an F-15 aircraft under a variety of operating conditions.

responses to white noise generated by a GenRad noise generator (Model #1381). The measurements were made in a reverberant room to ensure uniform incidence from all directions. We measured the responses of the three-microphone array, the ElectroVoice EV 985, and the Vought second-order gradient microphone, using the Racal Store 7-DS FM tape recorder. In addition to recording the sensor outputs, we observed the absolute sound pressure level (re 0.0002 microbar) as transduced by each sensor in a one-third octave band centered at 250 Hz. We used the 250 Hz band measurements as reference in determining sensor responses under aircraft cockpit noise.

For each sensor, we digitized the tape-recorded noise and computed the long-term average noise spectrum. Because the spectra we computed are DFT spectra (with uniform frequency bands) and the cockpit noise spectrum was measured in one-third octave (non-uniform) bands, we "tilted" the cockpit spectrum around the 250 Hz point to produce a uniform-band spectrum. We performed the tilting by subtracting $10 \log_{10}$ (center frequency/250) from the dB amplitude at the center frequency of each one-third octave band in the 0-5000 Hz range and linearly interpolating between the center frequencies. This has the effect of tilting the spectrum down by 3 dB/octave for frequencies above 250 Hz and tilting it up by 3 dB/octave for frequencies below 250 Hz. This step provides the required relationship between a one-third octave-band and a uniform-band

spectrum, where the latter has bandwidths all equal to the width (57 Hz) of the one-third octave band centered at 250 Hz (223-280 Hz).

Given this spectral representation of the cockpit noise, we wanted to estimate the response of the pressure microphone (M1) to this noise. Starting with the long-term average spectrum of the response of M1 under white noise, we obtained its absolute level by adding a constant to every spectral value; this constant was just the difference in dB between the measured 250-Hz (one-third octave band) level and the level at 250 Hz in the DFT spectrum. Then we obtained the estimated cockpit noise spe trum for M1 by adding to each point of the DFT spectrum the quantity $[C(i)-S_a(i)]$, where $C(i)$ is the ith amplitude in dB in the uniform-band cockpit-noise spectrum and $S_a(i)$ the ith amplitude in dB in the M1 spectrum adjusted to its absolute level as measured in the reverberant room.

For each of the other sensors we obtained its estimated cockpit noise spectrum by computing the following expression (using M1-M2 as an example):

$$S_c^{M1-M2} = S_c^{M1}(i)+[S^{M1-M2}(i)-S^{M1}(i)], \quad (10)$$

where $S_c^{M1-M2}(i)$ is the estimated ith cockpit spectral point for

M1-M2, S^{M1} is the ith cockpit spectral point for M1, and $S^{c}_{M1-M2}(i)$ and $S_{M1}(i)$ are the ith spectral points from the original unmodified spectra of the responses of M1-M2 and M1, respectively, under white noise.

Using the thecretically computed cockpit noise spectra, we examined briefly sensor performance in noise. Figs. 19 and 20 show plots of the cockpit noise spectrum, RMS speech spectrum, and peak speech spectrum for the sensors M1, M1-M2, and Vought and for the two positions 3 and 7. The RMS speech spectrum shown in the figures is the computed value of the long-term average spectrum plus 6 dB, to simulate the effect of a raised voice. (In our speech measurements, the speaker used a conversational speaking level; a raised voice RMS spectrum is approximately 6 dB above this level.) The peak speech spectrum was obtained by adding 12 dB to the RMS speech spectrum. The RMS and peak spectra were considered because of their relevance to the articulation index computation (see Section 4.8). From Figs. 19 and 20, we see that the speech peaks are significantly above the noise for the 0-3 kHz range for both gradient microphones (except for the second-order gradient microphone for position 3, where the range is only 0-2 kHz); however, the pressure microphone performance is substantially poorer. For the far positions (not shown in the figures), the pressure microphone is totally submerged in noise, and while the performance of the gradient microphones is degraded, it is noticeably better than the performance of M1.

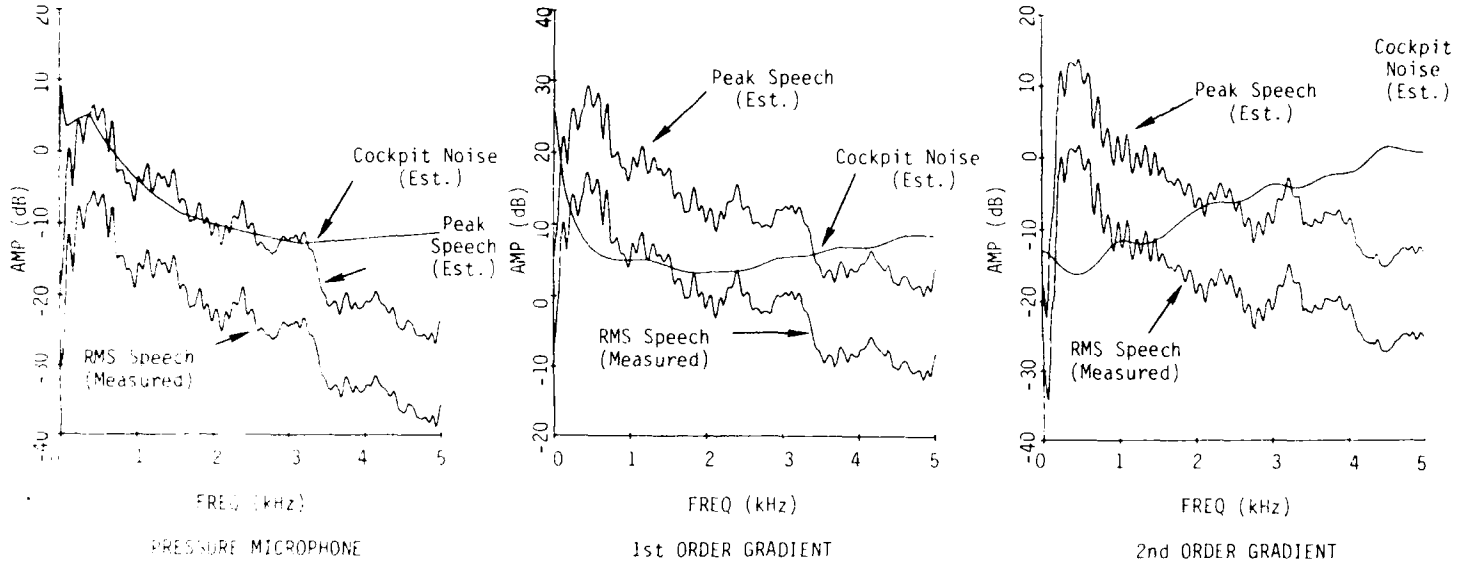


FIG. 19. Perfomance of three microphones (Position 3) in cockpit noise.

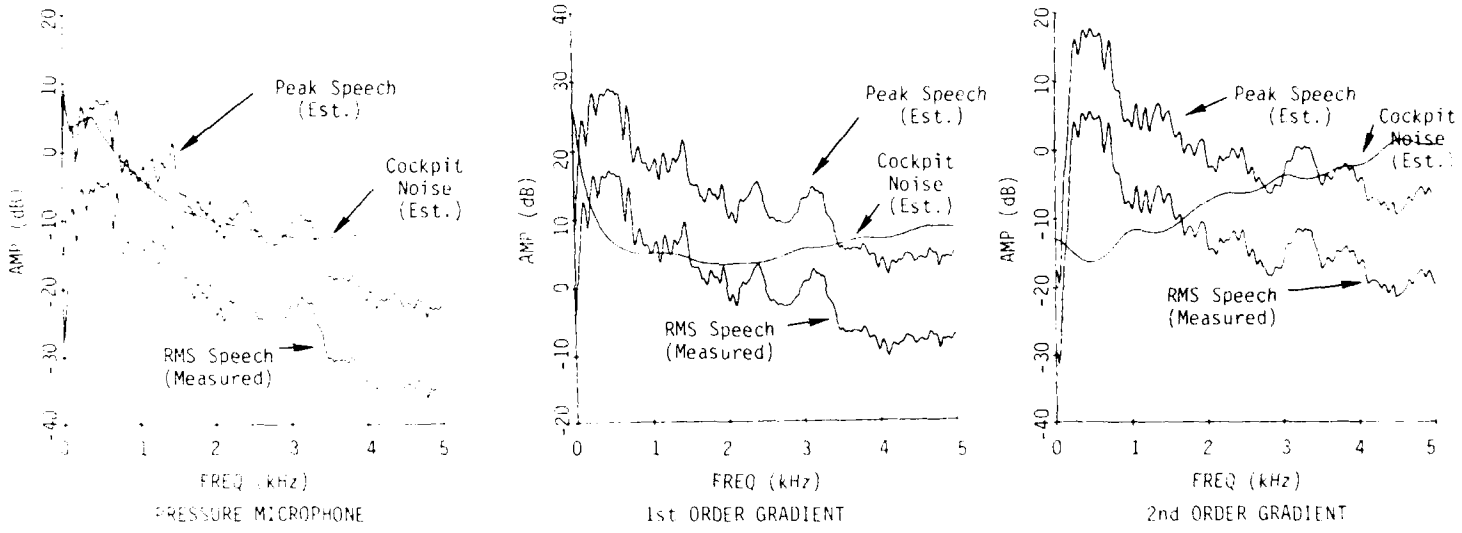


FIG. 20. Performance of three microphones (Position 7) in cockpit noise.

4.8 Comparison Using the Articulation Index

The AI estimates the intelligibility of speech in noise by the following method [9]. The dynamic range of speech is assumed to be 30 dB -- peaks are 12 dB above the RMS spectrum and the weakest sounds are 18 dB below. The spectrum is divided into 20 bands of non-uniform widths. The width of each band is empirically determined so that all bands contribute equally to the AI in the absence of noise. For each band we compute

$$W_i = \text{Min}\{(S_i - N_i), 1\}, \qquad (11)$$

where W_i is the ith band contribution to the AI, S_i is the average of the peak speech power in the ith band, and N_i is the average noise power in the ith band. W_i therefore represents the fraction of the total speech dynamic range that is above the noise in a given band, clipped to 1 if the entire 30 dB range is above the noise. If $S_i \geq 95$ dB (the pain threshold), S_i is set to 95; if $S_i \leq H_i$ (hearing threshold in band i), $S_i - N_i$ is set to 0. Finally, the overall articulation index is computed from:

$$AI = 0.05(W_1 + W_2 + \ldots + W_{20}). \qquad (12)$$

We developed and tested a program for computing the AI from the long-term average spectra of speech and noise. A raised voice condition was assumed in the AI computation.

We computed the AI scores for the eleven positions of each of the four sensors: M1, M1-M2, EV 985, and Vought. Fig. 21

shows plots of the AI scores averaged over near positions (1, 3, 7, and 9) and over far positions (2, 4, 8, and 10). As we did for the average spectral data in Section 4.5, we excluded positions 5, 6, and 11 from these averages. Clearly, source-to-sensor distance has a significant effect on the AI score. For all sensors, the near positions produce substantially higher AI scores than the far positions do. Notice, however, that the AI score for the pressure microphone (M1) at the near positions is far below that of the gradient microphones located at twice the distance. The low AI scores of M1 mean that speech transduced by M1 will not be intelligible.

It is not completely clear from the AI results how intelligible speech in noise from the other sensors would be. A rule of thumb states that an AI score above 0.4 leads to intelligible speech. None of the first-order gradient microphone positions scored that well. However, two positions of the Vought second-order microphone (7 and 1) scored above 0.4. Table 3 gives the AI scores for the four best locations of each sensor.

From these results, we conclude that positions 7 and 1 for the Vought microphone should be intelligible in aircraft cockpit noise at the 107 dB level used in this experiment. Position 7 for M1-M2 may be marginally intelligible. Other sensor/location combinations are less likely to produce intelligible speech in noise. This evidence supports the rationale for a multisensor configuration, since even the best single sensor performance

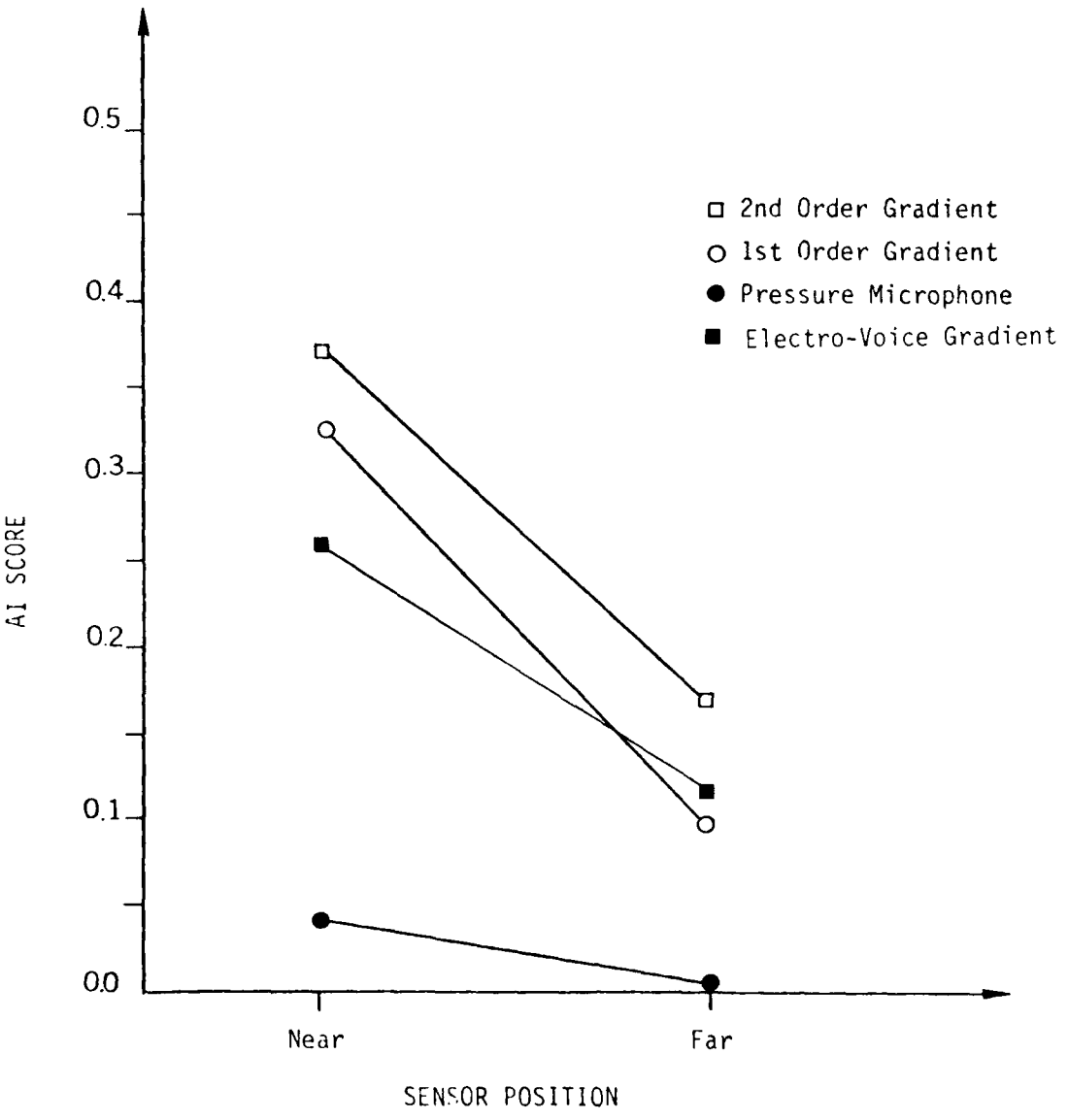


FIG. 21. The AI scores for various microphones.

Rank	M1		M1-M2		EV 985		Vought	
	Pos.	Score	Pos.	Score	Pos.	Score	Pos.	Score
1	7	0.058	7	0.363	7	0.271	7	0.437
2	5	0.043	3	0.347	3	0.258	1	0.407
3	9	0.043	9	0.329	9	0.246	3	0.339
4	3	0.030	1	0.260	1	0.239	9	0.316

TABLE 3. AI scores for the four best locations of the different sensors.

barely meets the minimum intelligibility requirement. Observing that position 7 consistently performs best for all sensors (according to the AI measure), we recommend it as the preferred location for any of the pressure and pressure-gradient microphones.

5. A TWO-SENSOR CONFIGURATION

From a review of the literature, we have identified at least four types of applications in which two or more sensors have been used for transducing the speech signal:

1. Two-microphone input for adaptive noise cancelling [11, 12];
2. Three-sensor input for speech training of the deaf [13];
3. Multimicrophone signal processing to remove room reverberation [14];
4. Multimicrophone signal processing for cocktail party effect [15];

In the applications 1, 3, and 4, individual sensor signals are combined to produce a single speech output. The second application uses the three sensor signals (a microphone near the mouth and two accelerometers, one on the throat and the other on the nose) as parallel inputs for a training system that displays the speech waveform, pitch period, and nasal accelerometer output. We note also that the third and fourth applications are not directly relevant to the purposes of this project since they involve microphones all located relatively far from the talker. Below, we describe our work on a new two-sensor configuration involving the accelerometer and one of the gradient microphones.

We noted above that the accelerometer signal has useful information in the low-frequency region extending up to about

2000 Hz. The attractive property of the accelerometer is that it is essentially insensitive to environmental noise. Therefore, we investigated a two-sensor configuration involving the accelerometer (located at position 6, see Fig. 16) and a gradient microphone (located at the near position 7). The output of this configuration is the sum of a lowpass filtered version of the accelerometer signal and a highpass-filtered version of the gradient microphone signal. The highpass filtering should substantially reduce the puff noise (as well as low-frequency environmental noise) picked up by the microphone. Before the two filtered signals are added, their relative overall levels must be adjusted to minimize any distortions in the combined signal.

To gain some insights about mixing the two sensor signals, we conducted several simple tests in a quiet environment. In the test setup, we amplified the M1-M2 output and the accelerometer signal independently using two amplifiers; used two variable cutoff filters for lowpass filtering the accelerometer signal and highpass filtering the signal from M1-M2; and mixed the two filtered signals using an audio mixer. We set the mixer channels to full scale; the relative levels of the signals being mixed were controlled with the amplifiers.

We explored three parameters in these tests: the highpass filter cutoff frequency for M1-M2, the lowpass filter cutoff frequency for the accelerometer, and the relative gain of the two sensor signals. Initially, we made the two cutoff frequencies

equal and varied this common frequency in 100 Hz increments from 500 Hz to 1200 Hz. For each value of the cutoff frequency, we kept the M1-M2 gain fixed at 12 dB and adjusted the accelerometer gain over the range 40-50 dB to obtain the most natural-sounding mix. Subsequently, we held the M1-M2 filter cutoff constant at 1200 Hz and varied the accelerometer filter cutoff frequency. During these tests, a male subject spoke continuously while a listener seated in another room made the cutoff frequency and gain adjustments and monitored the quality of the speech he heard over a headset.

We made several conclusions from these tests. First, for the common cutoff frequency case, we found that as the frequency was decreased, larger gain on the accelerometer signal was required to produce the most natural-sounding speech. The required accelerometer gain was about 6 dB higher at 500 Hz than at 1200 Hz. Second, a similar result was obtained for the case where we fixed the M1-M2 cutoff frequency at 1200 Hz and varied the accelerometer cutoff frequency; the needed gain adjustment (0 to 4 dB) was less than for the common cutoff frequency case. As an extreme case, we mixed the unfiltered accelerometer signal with the 1200-Hz highpass filtered M1-M2 signal. The mixed signal sounded as good as the signal in other cases. The use of the unfiltered accelerometer signal is particularly attractive in high-noise situations. Third, we observed that the intelligibility and naturalness of the mixed speech were not very

sensitive to moderate variations from the best mix settings of cutoff frequency and gain. Also, the best mix with the 1200 Hz crossover frequency did not sound significantly different from the best mix at 500 Hz. Speech in both cases sounded quite natural, although not as good as the M1-M2 output. The primary distortion was a boominess caused by the accelerometer, which we subsequently removed by highpass filtering the accelerometer signal at 200 Hz.

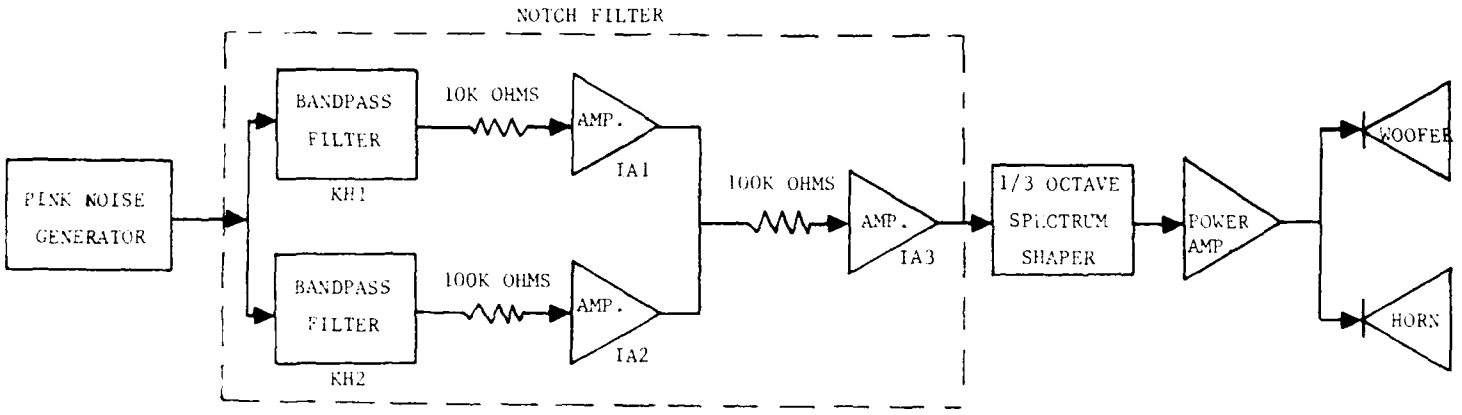
From these tests, we determined the range of parameter settings for achieving the best mixed signal in the quiet environment, with the expectation that these settings would have to be readjusted for the cockpit noise environment. For example, as the overall noise level is increased, the gain of the accelerometer signal may have to be increased further, as we want to reduce the noise level in the mixed signal. We also expect that as the noise level is increased, the combined output from the two-sensor configuration will be progressively better than the output from the gradient microphone. To test if this is in fact the case, we ran speech intelligibility and quality tests in a simulated cockpit noise environment, as described in the next section.

6. INTELLIGIBILITY AND QUALITY EVALUATION IN SIMULATED COCKPIT NOISE

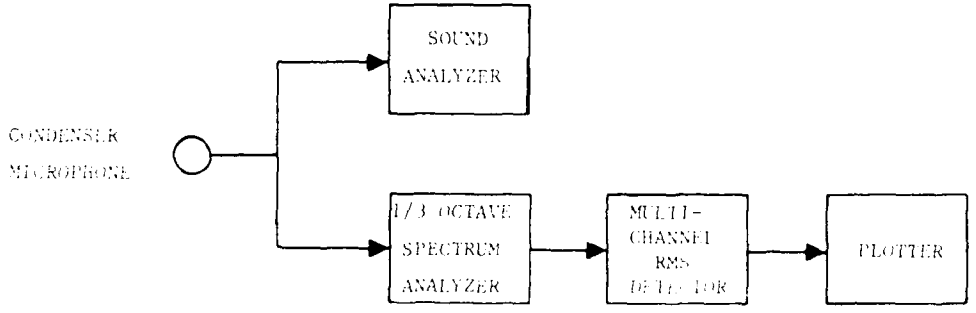
To determine if the two-sensor speech signal is better in noise than the single sensor signal, we ran speech intelligibility and quality tests in simulated cockpit noise at each of three levels. As we mentioned before, the overall level of the cockpit noise spectrum in Fig. 18 is 107 dB (re 0.0002 microbar). The other two levels we chose are 100 dB and 114 dB. Below, we describe how we simulated the cockpit noise and generated the test tapes, and present the results of the intelligibility and quality tests.

6.1 Simulation of Cockpit Noise

The tests were conducted in a reverberant room to ensure a diffuse, stable noise field. Fig. 22 shows block diagrams of the setup we used for generating and monitoring the noise, and Table 4 lists the equipment used. Referring to Fig. 22(a), we simulated the F-15 cockpit noise by shaping the spectrum of the output of a pink-noise generator with a one-third octave spectrum analyzer/shaper, amplifying the shaped noise to obtain the desired overall level, and playing the noise through a system of speakers (woofer and horn) that were located inside the reverberant room. The role of the notch filter inserted between the pink-noise generator and the one-third octave filter bank is discussed below.



(a) Noise generation



(b) Noise monitoring

FIG. 22. Block diagrams of the setup used for generating and monitoring the cockpit noise.

The noise monitoring setup, shown in Fig. 22(b), consists of the following items: a standard condenser pressure microphone (omnidirectional) suspended in the center of the reverberant room (near the place where the subject (talker) was located), to

Name and Model Number	Description
Ivie Electronic IE-20B	Pink/white noise generator
Krohn-Hite 3202	Variable cutoff lowpass/highpass filters
Ithaco 453M103	Calibrated amplifier
Genrad 1921	Real-time one-third octave spectrum analyzer/shaper
Crown M600	Power amplifier
Altec woofer	Low-frequency speaker system
University BP-12	Horn array speaker system
Bruel and Kjaer 4134	Condenser microphone
GenRad 1564	Sound and vibration analyzer
GenRad	Multichannel RMS detector
Hewlett-Packard 7015B	x-y plotter

TABLE 4. A list of the equipment used in generating and monitoring the cockpit noise.

measure the noise field; a sound and vibration analyzer to measure the overall (full spectrum) level of the microphone output; a one-third octave analyzer and a multichannel RMS detector to measure the spectrum of the microphone output in

one-third octave bands (by integrating over a period of about 16 seconds to ensure a stable measurement); and an x-y plotter to plot the measured noise spectrum.

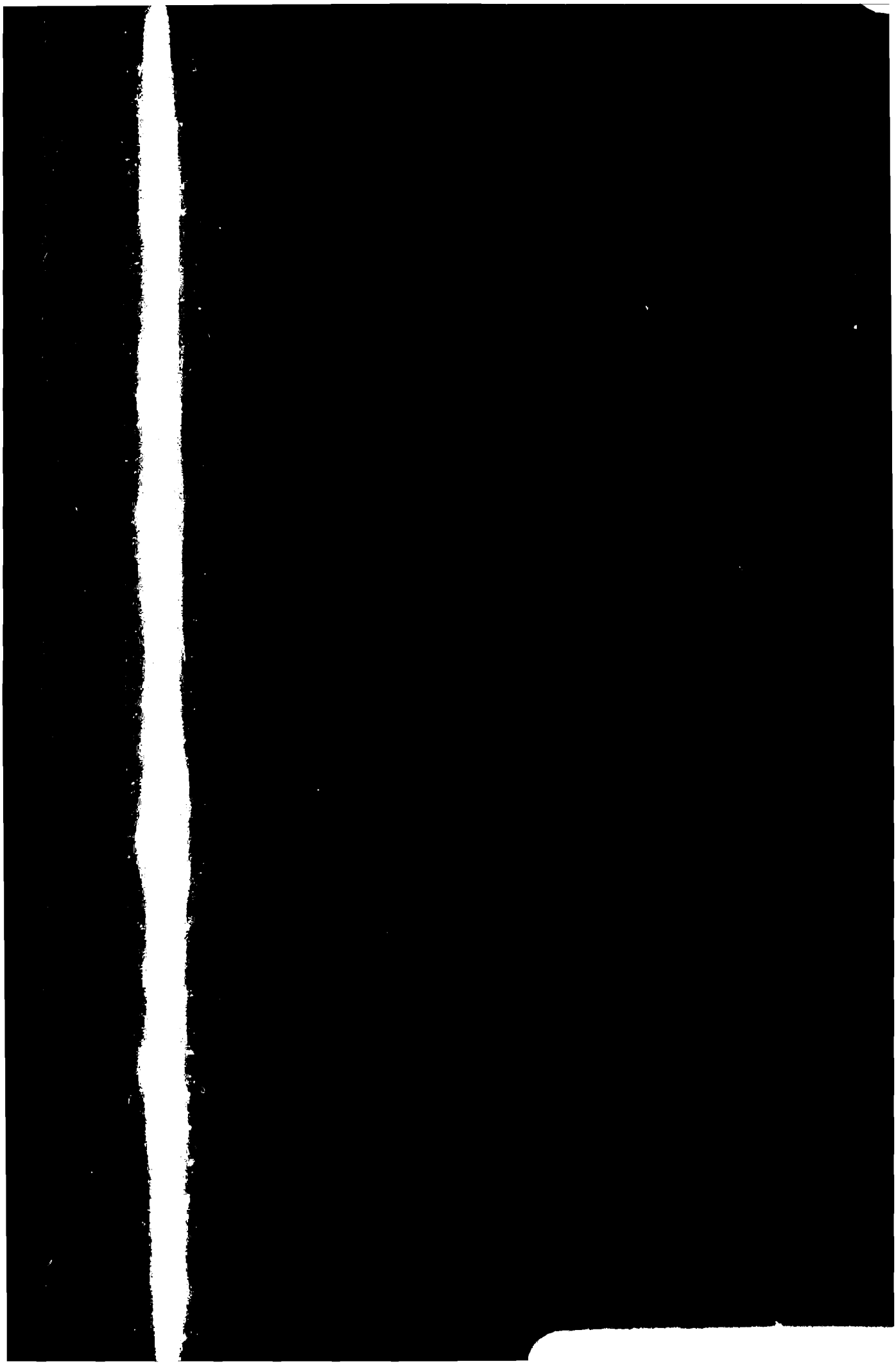
With the above setup, we adjusted the relative levels of the one-third octave bands of spectrum shaper iteratively. During these adjustments, we initially observed a pronounced peak in the measured noise spectrum in the 1500-2000 Hz range. To remove this peak and bring this part of the spectrum within an acceptably small deviation from the desired spectrum, we implemented and used an ad-hoc notch filter shown in Fig. 22(a). The amplifier IA3 was used to obtain the sum of the two bandpass filtered versions of the pink noise. A trial-and-error procedure was used to obtain the following settings: 10-1000 Hz bandpass for KH1, 0.1-100 kHz bandpass for KH2, -10 dB attenuation for IA1, 9 dB gain for IA2, and 7 dB gain for IA3.

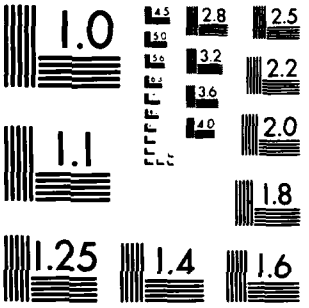
Figure 23 shows plots of the desired cockpit noise spectrum (with vertical bars indicating two standard deviations around the average levels) and the measured spectrum of the simulated noise. Except for frequencies above 5 kHz, the simulated noise spectrum was well within a standard deviation from the desired spectrum. We believe that the differences between the two spectra could have been reduced further if the one-third octave spectrum shaper had a larger dynamic range in some frequency bands and if the speaker system we used did not have certain unevenness in the frequency response. We feel, however, that the simulated noise was quite adequate for our purpose.

AD-A140 894 MULTISENSOR SPEECH INPUT(U) BOLT BERANEK AND NEWMAN INC 2/2
CAMBRIDGE MA V R VISWANATHAN ET AL. DEC 83
RADC-TR-83-274 F30602-82-C-0064

UNCLASSIFIED F/G 17/1 NL

END
DATE FILMED
6-84
DTIC





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

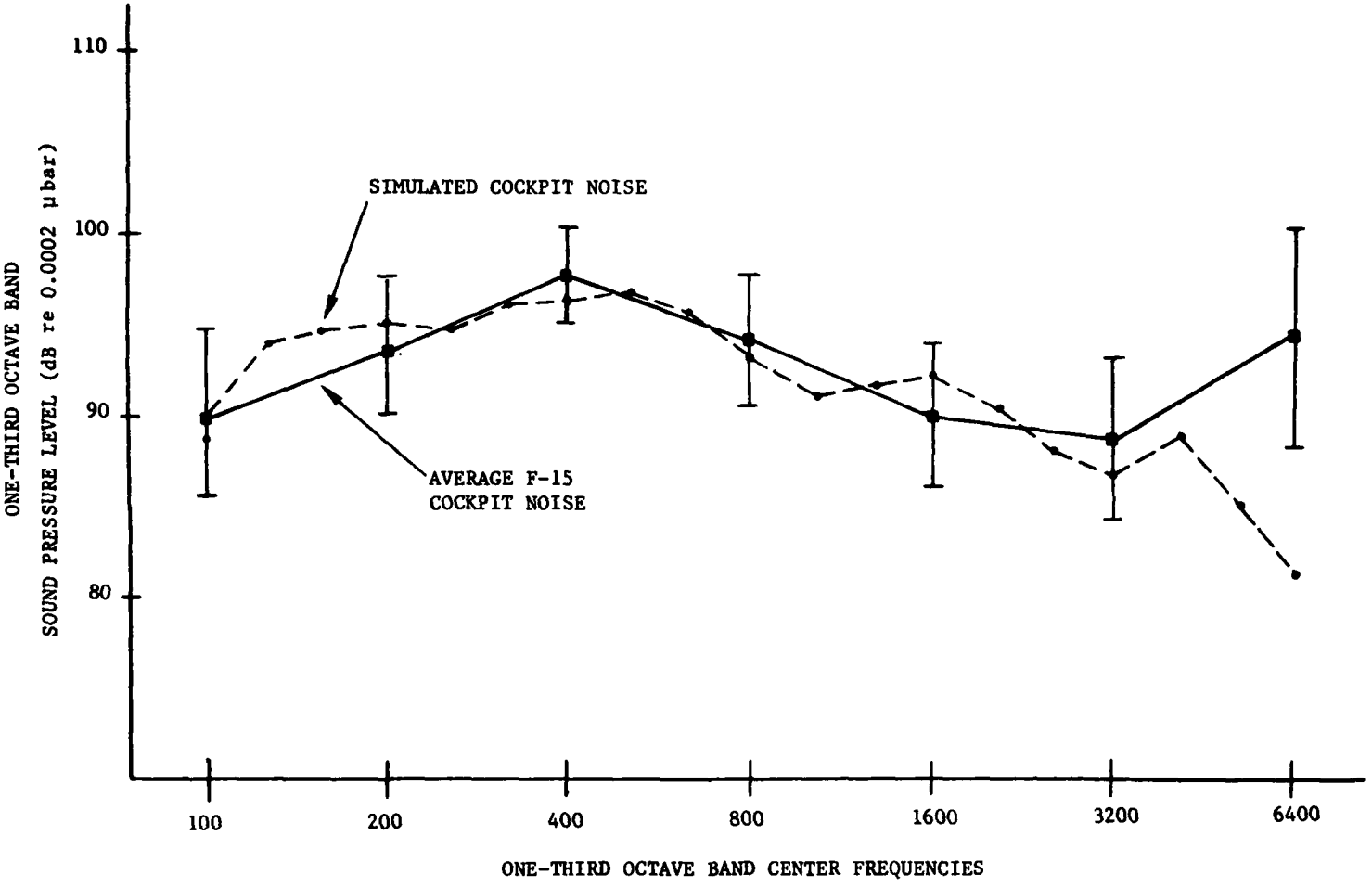


FIG. 23. The spectra of the desired and simulated F-15 cockpit noise.

Finally, we verified that the shape of the noise spectrum we generated did in fact remain constant (within $\pm$1 dB in each one-third octave band) as we changed the overall level from 100 dB to 107 and 114 dB. We also verified, by measuring the noise spectrum at several times during the experiment, that the shape of the spectrum did not drift over the course of several hours.

6.2 Generation of Test Tapes

We decided to use the diagnostic rhyme test (DRT) [16] for speech intelligibility evaluation and a simple 10-point rating test (see Section 6.3) for speech quality evaluation. We generated test tapes of speech materials from one male and one female talker under three different noise conditions (100, 107, and 114 dB) and in three sessions per talker. Each session, the talker used the accelerometer and a different one of the three gradient microphones (three-microphone array, EV 985, and Vought); this allowed us to record the combined two-sensor signal for each gradient microphone. Thus, we recorded a total of eighteen test runs (3 noise conditions x 2 talkers x 3 gradient microphones). For each run, we recorded the combined signal as well as the unfiltered gradient microphone and accelerometer signals.

For each test run, we used the following three sets of speech materials: 1) a randomized version of the 232-word DRT list; 2) the six sentences we used in our earlier noise-free

measurements (see Table 1); and 3) the ten isolated digits (zero through nine), spoken twice. For each of the 9 test runs per talker, we used a different randomized DRT list. We included the six sentences for the quality test and the digits for use in future testing of isolated word recognition systems.

Figure 24 shows the setup we used for generating the test tapes. We used two stereo tape recorders so that we could record the individual sensor signals as well as the combined two-sensor signal. For each test run, the individual sensor signals (accelerometer and gradient microphone) were recorded in stereo on recorder #1. On recorder #2 we recorded the gradient microphone signal and the combined signal in stereo (except in the case of the three-microphone array in which we recorded the output of M1 and the combined signal). We recorded the individual sensor signals on a single tape so that these synchronized signals could be combined in the future using techniques that are more effective than the simple method we used in the present study.

The method we used to combine the two-sensor signals was based on the results of our earlier experiment in a quiet room, which we reported in Section 5. The setup in this case was essentially the same as before: We amplified the gradient microphone and the accelerometer signals independently, using two (Ithaco) amplifiers; highpass filtered the two signals using variable cutoff (Krohn-Hite) filters; and mixed the two filtered signals using an audio mixer (Radio Shack).

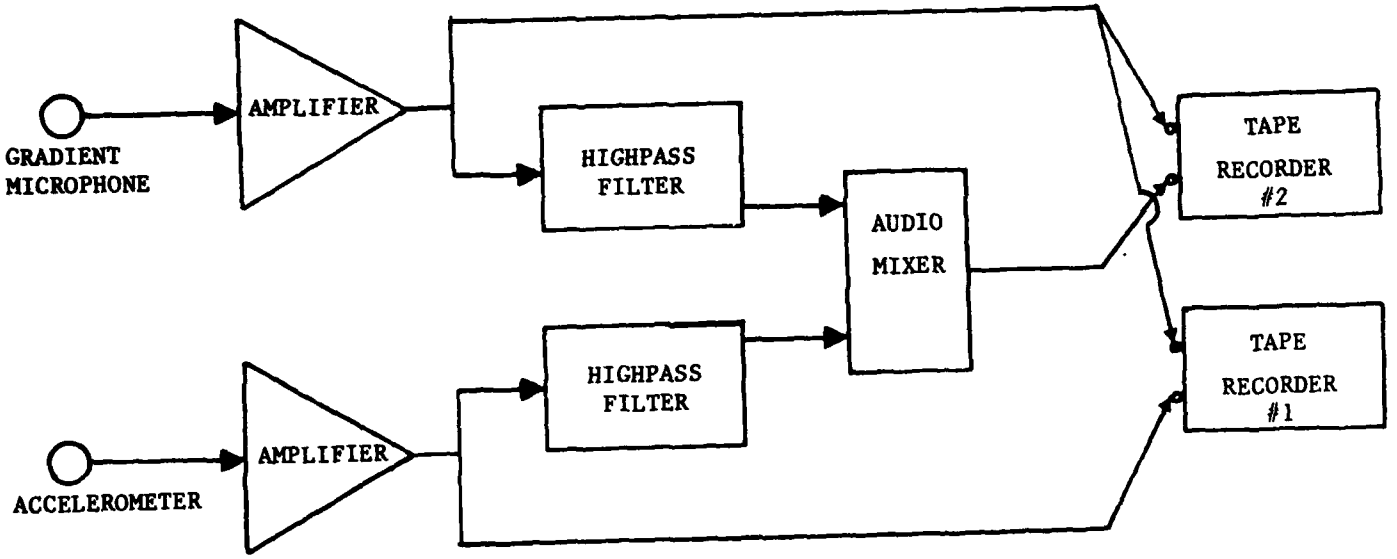


FIG. 24. The setup used for generating the intelligibility and quality test tapes.

Since the helmet (worn by the subject) to which the microphone boom was attached did not provide sufficient protection of the subject's ears from noise, both subjects used ear plugs. Before recording the test data, we adjusted the amplifier gain and the cutoff frequency for each sensor signal as a subject spoke the six sentences repeatedly inside the noisy reverberant room. As we expected (see Section 5), we found that retaining the high frequencies of the accelerometer signal (i.e., not lowpass filtering it) was beneficial in the noisy environment, and that it did not significantly degrade the naturalness of the speech. However, highpass filtering the accelerometer at 200 Hz helped to remove an unnatural boominess in the signal. As before, we found that the quality of the speech was not very sensitive to adjustments in the highpass cutoff frequency for the gradient microphone signal in the 500-1200 Hz range. Through informal judgments of speech quality and intelligibility of the two-sensor signal in all three noise conditions, we chose a cutoff frequency of 600 Hz.

Having found an initially satisfying mix, we made fine adjustments of the relative levels of the accelerometer and gradient microphone signals for each gradient microphone and each noise condition. With the amplifier gains fixed at their initially chosen values, we kept the accelerometer level control (on the audio mixer) at full scale and adjusted the gradient microphone level to obtain the best mix. The adjustments were

made as the subject repeated the six sentences until the listener was satisfied that the mix was subjectively optimal. Optimality in this case is rather loosely defined; we found it to be a compromise between the two conflicting objectives of less noise (which requires a lower gain for the gradient microphone signal) and better high-frequency information (which requires a higher gain for the gradient microphone signal). The best gain for the gradient microphone signal was found to be inversely related to the noise level.

Using the above-described procedure, we generated the test tapes for the eighteen conditions. Although we originally intended to locate each of the gradient microphones at position 7, in the actual test runs we located them inadvertently at position 3. For the accelerometer, we used position 6 (see Fig. 16). We instructed the talkers to read the DRT words at a pace of one word every 1.5 seconds and at a constant vocal effort, and the sentences in a natural manner.

From informal listening tests, we ascertained that the M1 output was very noisy and had poor intelligibility even at the 100 dB noise condition. The two-sensor signal had higher speech intelligibility and clearly better speech quality than the accelerometer signal, even at the 114 dB noise condition. Therefore, we did not include the M1 output and the accelerometer signal in the formal subjective tests reported below.

6.3 Results of Speech Intelligibility Tests

We ran the two-speaker DRT tests for M1-M2, Vought, and the corresponding two two-sensor configurations. (We did not include the data from EV 985, to limit the size of the testing effort.) We used five subjects, 3 males and 2 females, all college undergraduates. We played one of the unused test tapes (EV 985, 100 dB noise, female talker) as the practice run for the subjects.

Of the 232 words used, we scored only 192, with the rest being "filler words" [16]. The DRT scores averaged over the five subjects are listed in Table 5 for the various cases. Comparing each gradient microphone with the corresponding two-sensor configuration, we find that the 2-speaker average DRT scores are about the same at 100 dB, 6 to 7 points higher for the two-sensor case at 107 dB, and 10 to 15 points higher for the two-sensor case at 114 dB. The lower DRT scores for the talker BF (female) might be partly due to her soft voice. From the scores, we also see that the two-sensor configuration with M1-M2 is slightly better than the one with Vought.

6.4 Results of Speech Quality Tests

For speech quality evaluation, we used the gradient microphone signal and the corresponding two-sensor signal, for each of the eighteen test conditions. From the resulting

Noise Level →		100 dB		107 dB		114 dB	
Microphone Used	Talker ID	Microphone	Two-Sensor Config.	Microphone	Two-Sensor Config.	Microphone	Two-Sensor Config.
BBN Array (M1-M2)	KK (Male)	92.9	95.8	91.2	92.3	84.4	90.2
	BF (Female)	87.9	87.3	70.6	83.1	59.2	73.5
	Average	90.4	91.6	80.9	87.7	[illegible]	81.9
Vought	KK (Male)	91.4	90.0	87.9	89.2	86.0	89.8
	BF (Female)	80.5	81.4	54.6	65.5	36.0	63.3
	Average	86.0	85.7	71.3	77.4	61.0	76.6

TABLE 5. DRT scores for two gradient microphones and two two-sensor configurations, under three noise levels.

thirty-six sets of 6-sentence stimuli, we prepared a single test tape as follows. The test tape begins with a practice section containing four 6-sentence sets illustrating the range of qualities over the ensemble of conditions under test. The remainder of the tape has the thirty-six 6-sentence sets included in a random order, followed by a repetition of the first five sets; the rating scores on the last five sets must be compared with the scores on the first five to evaluate the consistency of the listener.

In our speech quality tests, subjects listened to the six sentences as a set and rated the quality on a scale of 1 to 10, with 10 being the best and 1 being the worst. Subjects were instructed to use the full rating scale. We initially used five subjects, one male (college undergraduate) and four females (3 high school students and 1 BBN employee), none of whom was experienced in the rating task. As we examined the rating scores, we observed a rather large variability, with one of the subjects being quite inconsistent. We decided to exclude the scores of this subject from further analysis. Also, we reran the quality test, using 3 staff members from our speech processing department. These three have had extensive experience listening to speech from various speech coding and synthesis systems, but were otherwise not familiar with the specific conditions being evaluated in this test. In Table 6, we give the quality ratings averaged over the 4 naive or inexperienced listeners (with the

fifth one being excluded), over the 3 experienced listeners, and over all 7 listeners, for each of the various test conditions. The scores given in parentheses are those computed from the ratings of the corresponding sets repeated at the end of the test tape. From the results given in Table 6 it is clear that the two-sensor signal has a higher quality than the gradient microphone signal that was used to derive the two-sensor signal. The quality difference is about 1.5 to 3 points for all noise conditions.

Noise Level (dB)	Male Talker (KK)						Female Talker (BF)					
	Microphone			Two-Sensor Configuration			Microphone			Two-Sensor Configuration		
	Naive	Experienced	All	Naive	Experienced	All	Naive	Experienced	All	Naive	Experienced	All
BBN Array (M1-M2)												
100	9.0	8.3	8.7	9.8	10.0	9.9	4.5	3.7	4.1	7.8	5.7	6.9
107	5.5	7.0	6.1	7.3	8.0	7.6	3.5 (4.0)	4.0 (2.3)	3.7 (3.3)	5.8	5.0	5.4
114	3.3	2.0	2.7	5.8	6.7	6.0	1.5	1.3	1.4	4.3	4.3	4.3
EV 985												
100	8.5 (8.0)	8.3 (8.0)	8.4 (8.0)	9.5	10.0	9.7	6.5	6.3	6.4	7.8	7.3	7.6
107	5.8	4.7	5.3	8.3	7.7	8.0	2.3	2.0	2.1	5.8	4.3	5.1
114	5.3	2.7	4.1	7.0	6.3	6.7	2.5 (2.3)	1.7 (1.7)	2.1 (1.9)	3.8	3.0	3.4
Vought												
100	8.8	8.0	8.4	7.3	9.7	8.3	6.0	3.3	4.9	7.5	5.0	5.6
107	7.0	6.3	6.7	7.5	8.7	8.0	3.3	1.0	2.3	5.0 (5.5)	3.3 (3.3)	4.3 (4.6)
114	5.5	6.0	5.7	8.5 (8.3)	8.3 (7.3)	8.4 (7.9)	1.8	1.0	1.4	2.3	3.3	2.7

TABLE 6. Mean speech quality ratings for three gradient microphones and three two-sensor configurations, under three noise levels.

7. SUMMARY AND FUTURE RESEARCH

In summary, we note three important contributions of this work. First, we measured the sound field during speech in the vicinity of the talker using pressure and pressure gradient microphones and the vibrations on the head and neck of the talker using an accelerometer. For these measurements, we used five talkers, a specially selected set of speech materials, and eleven selected positions for each sensor. Second, we investigated the properties of the measured signals from the various sensor types and positions through long-term and short-term spectral analyses and using articulation index scores computed assuming aircraft cockpit noise. As an important by-product of this investigation, we determined the best location for each of the various sensors. Third, we developed and tested several two-sensor configurations, each consisting of a gradient microphone and an accelerometer. The two-sensor signal is a weighted sum of filtered versions of the signals from the two individual sensors. The results we obtained from speech intelligibility and quality evaluation of several gradient microphones and two-sensor configurations in simulated aircraft cockpit noise snow clearly that each two-sensor signal outperforms its constituent gradient microphone signal and that the performance improvement generally increases with the noise level.

We suggest below five major areas for future research.

Short-term spectral analysis of the measured sound-fi data: Only a limited analysis was carried out in this work. more detailed and thorough analysis is required for understanc the frequency-dependent behavior of the signals from the vari sensor types and positions. We expect that the results from s analysis may lead to multisensor designs involving paral inputs.

Investigation of multisensor configurations: A sir method of combining the individual sensor signals, using high filters and based on informal listening, was employed in t work. Further work may be performed to develop more effect methods for combining the sensor signals and to develop multisensor designs involving parallel inputs.

Analysis of data from other talkers: The sound-field c measured from the other four talkers may be analyzed investigate the effects of speaker variability on the resu obtained in this work. (We used data from one male subject.)

Use of multisensor signals in speech recognition and sp coding: We need to evaluate the performance and robustnes speech recognition and speech coding systems with multise speech input.

New Speech Processing Methods: New speech processing methods (analysis, feature extraction, or template matching) may need to be developed to take full advantage of multisensor speech input, especially, the parallel-input version.

8. REFERENCES

1. A.J. Brouns, "Development of Second-Order Gradient, Noise-Cancelling Microphone," Final Report, Contract N00039-80-C-0146, Vought Corporation Advanced Technology Center, March 1981.

2. J.L. Flanagan, "Analog Measurements of Sound Radiation from the Mouth," J. Acoust. Soc. Amer., Vol. 32, pp. 1613-1620, Dec. 1960.

3. G. von Bekesy, "Uber die Vibrationsempfindung," Akust. A., Vol. 4, pp. 316-334, 1939.

4. H.M. Moser and H.J. Oyer, "Relative Intensities of Sounds at Various Anatomical Locations of the Head and Neck during Phonation of the Vowels," J. Acoust. Soc. Amer., Vol. 30, pp. 275-277, April 1958.

5. K.N. Stevens, D.N. Kalikow, and T.R. Willemain, "A Miniature Accelerometer for Detecting Glottal Waveforms and Nasalization," J. Speech and Hearing Res., Vol. 18, pp.594-599, Sept. 1975.

6. J.H. Oyer, "Relative Intelligibility of Speech Recorded Simultaneously at the Ear and Mouth," J. Acoust. Soc. Amer., Vol. 27, pp. 1207-1212, Nov. 1955.

7. H. Ono, "Improvement and Evaluation of the Vibration Pick-up-type Ear Microphone and Two-Way Communication Device," J. Acoust. Soc. Amer., Vol. 62, pp. 760-768, Sept. 1977.

8. A.J. Brouns, "Experimental Wide-Bandwidth Tooth-Contact Microphone," J. Audio Eng. Soc., Vol. 19, pp. 41-45, Jan. 1971.

9. L.L. Beranek, Acoustics, McGraw-Hill, New York, 1954.

10. R.L. Miller, R.D. Bruce, F.N. Iacovino, and A.W.F. Huggins, "Simulation of Cockpit Noise Environments in Four Tactical Aircraft for the Purpose of Testing Speech Intelligibility," BBN Report 4465, Final Report, Subcontract for Lincoln Laboratory, MIT, Sept. 1981.

11. B.J. Widrow et al., "Adaptive Noise cancelling: Principles and Applications," Proc. IEEE, Vol. 63, pp. 1692-1719, Dec. 1975.

12. D. Pulsipher, S.F. Boll, C. Rushforth, and L. Timothy, "Reduction of Nonstationary Acoustic Noise in Speech Using LMS Adaptive Noise Cancelling," Proc. IEEE International Conf. Acoustics, Speech and Signal Processing, Washington, DC, pp. 204-207, April 2-4, 1979.

13. R.S. Nickerson, D.N. Kalikow, and K.N. Stevens, "Computer-Aided Speech Training for the Deaf," J. Speech and Hearing Disorders, Vol. 41, pp. 120-132, Feb. 1976.

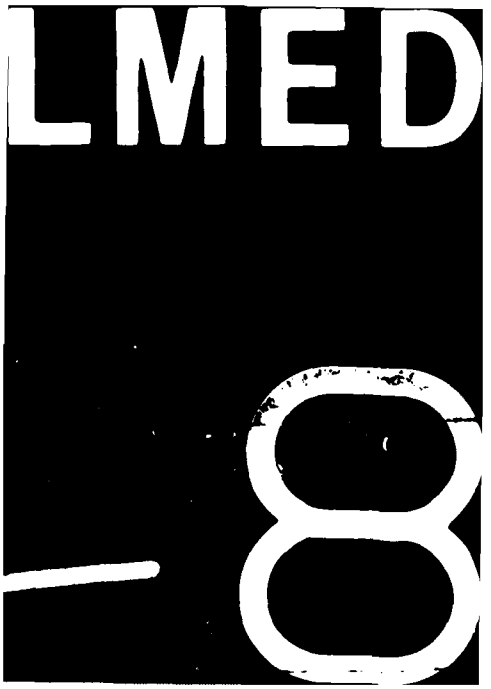
14. J.B. Allen, D.A. Berkley, and D. Blanert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals, "J. Acoust. Soc. Amer., Vol. 62, pp. 912-915, Oct. 1977.

15. O.M.M. Mitchell, C.A. Ross, and G.H. Yates, "Signal Processing for a Cocktail Party Effect," J. Acoust. Soc. Amer., Vol. 50, No. 2 (Part 2) pp. 656-660, 1971.

16. W.D. Voiers, A.D. Sharpley, and C.J. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," Final Report, Contract No. AF19628-70-C-0182, Air Force Cambridge Research Laboratories, 1973.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.


LMED
8